Proposal: Supporting Volunteer Moderation Practices in Online Communities

by

Joseph Seering

Submitted to the Human-Computer Interaction Institute
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Human-Computer Interaction

at Carnegie Mellon University

August 2019

Contents

Chapter 1

Motivations

1.1 Introduction

In May of 1978, the "CommuniTree #1" online Bulletin Board System (BBS) launched in the San Francisco Bay area. Built from the CommuniTree Group's idea to structure online conversation in threaded, tree-style structures based around core "conference" topics, it was the most successful entry into the very new space of online communities; while the first set of these virtual bulletin boards, developed in the mid-1970s, were searchable usually only either in alphabetical order or in the order messages were posted, CommuniTree #1's design allowed for conversations to move fluidly in different directions. CommuniTree #1 started off with a grand vision of the power of social technology – the first "conference" opened with the bold statement "We are as gods and might as well get good at it" [8, p. 5]. Per Allucquère Rosanne Stone's account, the participants (mostly academics and other researchers) saw themselves "not primarily as readers of bulletin boards or participants in a novel discourse but as agents of a new kind of social experiment" [8, p. 6]. In 1982, Apple entered into an agreement with the American government to provide schools with Apple computers as a substitute for paying taxes. This led to an influx of teenage, mostly male users into virtual spaces previously reserved for the intellectual elite. These students, upon discovering CommuniTree, flooded the board with obscene messages and crude jokes, an onslaught for which the existing users were completely unprepared. CommuniTree

had been launched with essentially no moderation features, and the students' incursions forced system operators to completely purge the system almost daily. Within a few months, CommuniTree was dead.

The online, self-governing utopia that was CommuniTree lasted for less than half a decade. Relying purely on the goodwill of its homogenous user-base, it had managed to survive and even thrive, but when confronted by a new set of users with different values and goals it collapsed. This represents the first major online instance of a failure in moderation, and it effectively ended the dream that the internet could function without at least some tools and strategies for maintaining order. There are spaces online today that are designed with little focus to moderation – Slack, for instance, has minimal moderation tools because it is designed for groups and teams with shared goals and existing standards for behavior – but the importance of moderation tools in virtually every public space online has become widely accepted.

The details involved in the development and use of moderation tools, however, are not at all simple. Extensive research over the past nearly forty years has identified problems that haven't yet been resolved by existing tools. Despite tremendous growth in the adoption of online social systems, now used by a strong majority of the Earth's population, online conflict is far from a solved problem and is perhaps a bigger problem than it ever has been.

There are a variety of factors that contribute to the depth of the problem. First, a number of persistent technical challenges have proven resistant to all attempts to address them. For example, the underlying structure of the internet makes it easy for users to adopt new identities in many cases where accounts are not directly tied to a fixed personal identifier. Users who are punished or removed from a particular social platform can frequently create a new account and continue misbehaving, albeit in some cases with a loss of accrued connections or content associated with their previous account. Friedman and Resnick call this the problem of "cheap pseudonyms" [2], and it has remained largely unsolved since the early days of online communities (see e.g., [7]).[1]

[1]Though the difficulty in tracking misbehaving users across multiple usernames causes problems

Another major technical challenge to moderation is handling the unimaginably massive volumes of content that are generated on modern social platforms. As I discuss later, there is orders of magnitude more content generated than teams of platform employees could ever have time to review, and the patchwork systems of algorithms and hired contractors that platforms currently employ have not proven successful in addressing serious social problems. While it is arguable that platforms on the very early social web had the capacity to vet and moderate all content posted to their sites, that era has long passed. This challenge is compounded by complex legal frameworks and debates about how and when platforms are obligated to moderate. Most notable among these are Section 230 of the Communications Decency Act of 1996 in the United States and the General Data Protection Regulation of 2018 in the European Union, which are polar opposites in the level of responsibility they place on platforms.

The primary social challenge in moderation, beyond the social implications of the aforementioned problem of scale, is the relative lack of social signals in online communication. Conversations in online communities simply do not have the depth of signals that face-to-face conversations have, from body language to facial expressions to tone and pacing of speech.[2] While virtual reality offers some interesting opportunities for increased fidelity in online communication, this technology is still far from widespread use in any level of fidelity rivaling real-world complexity of available social information. This phenomenon, explored in depth in the 1980s by Kiesler and colleagues (e.g., [4]) and later in the 1990s notably by Donath [1], can lead to misunderstanding and sometimes disinhibition and can facilitate deception.

Perhaps most important, however, are the philosophical issues that surround online moderation. It is not an exaggeration to say that the decisions regarding who has control over how the internet is moderated will shape the future of political speech and social interaction as a whole. While this has often been discussed as a conflict

for moderation, it is worth noting that this form of pseudo-anonymity also has an important role in protecting certain forms of speech.

[2]Voice chatrooms, as currently used on platforms such as Discord, allow for the latter two types of signals, but these chatrooms do not accommodate particularly large communities simultaneously.

between free expression and censorship, the reality is far more complex. As Gillespie [3] discusses at length, the current state of online moderation requires platforms to balance ethical principles with technical feasibility, political pressures, and legal responsibilities in making these decisions. Even beyond these internal corporate debates are societal debates about where the power to regulate should sit, whether at the level of platforms, politicians, or people.

Given the above, I believe that this is an appropriate moment in time to consider what the future of these social spaces can look like, balancing an understanding of technical feasibility with knowledge of behavioral principles. The state of current and predictable future technology is such that platforms will not realistically be able to address all of the pieces of content that need to be addressed in the depth that is necessary to make informed decisions. Based on this, I have begun to focus on the ability for volunteer user moderators to contribute more directly to these large-scale moderation processes. Ultimately, the users involved in a given social situation are in most cases better situated to understand the relevant social context than algorithms or contractors, and the prevalence of volunteer user moderators across multiple current platforms suggests that users are already reasonably motivated to care for their own spaces. If these efforts are supported and integrated into platforms' existing processes, the problems described above can all be made more manageable.

In this proposal I discuss the tools and strategies that are currently employed in online spaces and the sociotechnical systems in which they are embedded, the research that has studied the problems they address and their effectiveness, and the avenues for future research in improving them. I begin in Chapter 2 with a review of existing literature in moderation, separating it into three general categories and detailing these categories' perspectives. In Chapter 3 I present a framework for the processes of user-driven moderation from my interviews with 56 volunteer moderators from Facebook Groups, Reddit, and Twitch.[3] In Chapter 4 I present a quantitative analysis of the impact of imitation and deterrence effects on user behaviors, focusing

[3]This chapter is drawn from my 2019 paper, "Moderator Engagement and Community Development in the Age of Algorithms" [6].

on the potential for these effects to help support moderation practices. [4] Finally, in Chapter 5 I propose work I will complete in my dissertation to build upon this prior work in exploring the potential for improving moderation of online communities in the future.

Bibliography

[1] Judith Donath. Identity and Deception in the Virtual Community. In Marc A Smith and Peter Kollock, editors, *Communities in Cyberspace*, pages 27–58. Routledge, London and New York, 1st edition, 1999.

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

[3] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.

[4] S Kiesler, J Siegel, and T W McGuire. Social psychological aspects of computer-mediated communication. *American Psychologist, 39(10)*, pages 1123–1134, 1984.

[5] Joseph Seering, Robert E Kraut, and Laura Dabbish. Shaping Pro and Anti - Social Behavior on Twitch Through Moderation and Example - Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017.

[6] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society*, 2019.

[7] Anna DuVal Smith. Problems of Conflict Management in Virtual Communities. In Marc A Smith and P Kollock, editors, *Communities in Cyberspace*, pages 135–166. Routledge, London and New York, 1st edition, 1999.

[4] This chapter is drawn from my 2017 paper, "Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting" [5].

[8] Allucquère Rosanne Stone. Will the Real Body Please Stand Up? In Michael Benedikt, editor, *Cyberspace: First Steps*, pages 81–118. MIT Press, 1991.

Chapter 2

Prior work

2.1 Introduction

Though it is difficult to define exactly which research falls within the domain of online moderation, I focus in this chapter on research that identifies problematic social behaviors online, analyzes social or technical properties of these behaviors, studies responses to these behaviors, or considers new approaches to responding. I argue here that this research can be grouped into three primary categories, each of which approaches moderation from a different analytical perspective. The first of these is the *institutional* perspective, which is currently the most prominent in public discourse. It takes the perspective of the major social platforms, including Facebook, Twitter, Instagram, and various others, to consider what can be done about various problematic behaviors. The second perspective, the *behavioral* perspective, focuses specifically on properties of these behaviors and the characteristics of users who engage in them or are impacted by them. The third perspective, to which I aim to contribute with my research, focuses on properties of the social structures, e.g., communities, through which people interact online, with the goal of understanding how these structures. I argue that this third perspective, the *social systems* perspective, has been underrepresented in both public and academic discourses, and that it has much to contribute to the conversation in determining how these platforms are designed in the future.

In this chapter I first outline each of these three categories in depth and identify

major theoretical and empirical works in each. Each category can trace its origins to different fields of research, and I describe these influences in this chapter and how they lead to different perspectives and different framings of similar problems. I conclude by describing how my research fits into this space and I discuss the primary theoretical perspectives that underlie my work.

2.2 Platforms as governors; platforms as custodians

In his 2018 book *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*, Communications & Media scholar Tarleton Gillespie provides the defining institutional perspective of this era of moderation research [22]. Per the book's title, Gillespie argues that social media platforms are caretakers of the online world. They are custodians both in the sense that they must keep these platforms *clean* and in that they have *custody* of modern discourse. In his book, he discusses the broader challenges platforms face in balancing their general philosophies, ethics, technical feasibility, and legal requirements. He describes how these factors manifest in how they write community guidelines, how they choose which moderation features to employ, and how they structure their organizations to be able to review content and respond. These topics are the major focus areas within the institutional perspective in moderation research, and the scholars in this domain typically draw from fields that theorize organizations, media, and political processes, e.g., communications, law, and media studies.

One of the earliest works analyzing the sociopolitical position of platforms in 2010, aptly named "The Politics of 'Platforms" ', also comes from Gillespie [21]. This era saw a major transition the social structure of the internet from "online communities" to "platforms" and "social networks", changing both the structures of social relations online and the language used to describe them. Gillespie sees platforms as inherently political entities that have attempted to maintain an image of neutrality. He analyzes the underlying meaning of the word "platform", which had by that time become ubiquitous in describing Facebook, YouTube, and other rapidly-growing sites. He

argues that sites pitch themselves as platforms for a variety of reasons. For example, they self-present as *technical* platforms to highlight the value of their technology in facilitating future innovation. More controversially, they call themselves platforms for *speech*, evoking imagery of both an open, level playing-field and a space to elevate users' speech. It is this image of neutrality in the domain of speech that has come under question most in recent years, and the inability for platforms to be truly neutral is core to the thesis of Gillespie's work.

Though *Custodians* may be the most prominent modern work taking this institutional perspective, other scholars have argued similar points. Legal scholar Kate Klonick's "The new governors: The people, rules, and processes governing online speech" begins with the argument that social platforms must be understood as private systems of governance that sit between regulators and speakers [37]. She highlights the history of regulation of content moderation in the United States, notably including Section 230 of the Communications Decency Act of 1996, which specifies the requirements for platforms in moderating users' speech. Though originally intended to protect platforms that engaged in "good faith" content moderation, it has been interpreted by courts in ways that give platforms almost complete leeway in what and how they moderate[1]. Section 230 has been the focus of other legal scholars' work in topics related to moderation; Danielle Keats Citron's *Hate Crimes in Cyberspace* detailed the experiences of victims of the non-consensual distribution of sexual imagery or video (also known as "revenge porn") and the role of Section 230 in protecting platforms that host this type of content [9]. Klonick notes that, despite a general lack of legal requirement to do so, platforms do tend to actively moderate content. She argues that this is the result of several factors, including a philosophical basis in American free speech norms, a sense of corporate responsibility, and the necessity for platforms to provide an appealing space for social interactions. Ultimately, though Klonick and Gillespie differ somewhat in how they approach platforms' roles, both focus on moderation through the lens of the substantial social, technical, and

[1]The Electronic Frontier Foundation, writing from a strong cyber-libertarian perspective, calls Section 230 "The most important law protecting internet speech" `https://web.archive.org/web/20190710114401/https://www.eff.org/issues/cda230`

political power that modern platforms wield as it intersects the tremendous pressures they face.

While Klonick and Gillespie provide broad perspectives on the roles of platforms, describing them as *governors* and *custodians* respectively, other scholars have focused in depth on specific processes within these institutions. Pater et al. [53] examined various platform policy documents, from terms of service to community guidelines to parental and teen/youth guides, to understand how they define harassment. As of early 2016, none of the 15 major platforms they analyzed provided a specific definition for harassment in any of the 56 documents they collected, and only Twitter and Instagram provided descriptions of behaviors that were taken into consideration when determining whether actions would be defined as harassment.[2] In the three years since Pater et al.'s paper was published, platforms' policies have become increasingly broad if not more detailed. For example, on July 9th, 2019, Twitter released an amendment to its policy on "hateful conduct", adding religious affiliation to the list of categories protected from threats, harassment, or promotion of violence. This list had previously included race, ethnicity, national origin, sexual orientation, gender, gender identity, age, disability, and serious disease.[3] Though relatively few users read these policies in depth, the way that they are written is a manifestation of how platforms balance their competing motivations [22, p. 45], describing their philosophies as mediated by technical feasibility and legal pressures.

Other recent work has examined the impact these policies have on users as expressed through content removal. Myers West [82] studied users' reactions to content removal and folk theories about how those systems work, noting that users are often left to speculate about reasons for removal due to a general lack of transparency in moderation actions. She found that users attributed their ban or the removal of their content to a variety of sources, frequently suspecting that they had been reported by a "friend" or another user, but sometimes also suspecting a general bias on the

[2]These behaviors included "repeated unwanted contact" on Instagram and "reported behaviors [that are] one-sided or include threats".

[3]`https://web.archive.org/web/20190710012822/https://blog.twitter.com/official/en_us/topics/company/2019/hatefulconductupdate.html`

part of the platform employees. A number of participants also reported that their bans had significant negative impacts on their lives, including disconnecting them from family and loved ones or hindering or blocking their careers.[4] Suzor et al. built on this work by identifying specific ways in which increased transparency could help educate users and establish a sense of trust in these processes [75]. Achieving these levels of transparency is a challenge; increased transparency would expose internal processes to more public debate, but could also highlight the many ways in which these companies are currently ill-equipped to make context-sensitive decisions.

Finally, other recent work has examined the socio-technical methods through which platforms actually make moderation decisions. Grimmelmann provides an initial taxonomy of these methods from the perspective of specific actions platforms can take [23]. These include "excluding", "organizing", "princing", and "norm-setting". Grimmelmann also provides a taxonomy of the different ways in which each of these can be performed, comparing, e.g., centralized and decentralized moderation, automatic and manual moderation, and ex post and ex ante moderation. Subsequent work has examined the both the commercial realities and driving philosophies of implementation of these and other approaches. Crawford and Gillespie detailed the concept of a "flag" on social media, the tool through which users report content that they believe violates rules or is inappropriate for the platform [11]. The flag is core to moderation processes on most major social platforms; given the enormous volume of content that is produced, companies often must rely partially on users' reports to identify which content to examine. However, flags can be designed in a variety of ways. Platforms may require users to specify which of many reasons they are reporting a piece of content for, effectively defining what is and is not permitted by limiting what can and cannot be reported.[5] Blackwell et al., though not working from an institutional perspective, detail the consequences of this sort of classification – while it can validate

[4]The latter case notably included users who ran or advertised their small businesses on social media platforms like Facebook or Instagram.

[5]For example, at the time of writing, YouTube offers nine categories for users to pick from when reporting a piece of content. These range from "Sexual Content" to "Child Abuse" to "Promotes Terrorism". These categories have between zero and six subcategories to choose from. There is no top-level "Other" reason for reporting.

users' experiences by making norms clear, it can also invalidate the experiences of users whose experiences do not match the classification scheme [4]. Flagging mechanisms can also be gamed or abused; Klonick and Gillespie note cases where organized groups of users flagged content *en masse* as a form of attack on the content creators [11, p. 420-421]. Despite these vulnerabilities, most major social platforms rely on flagging systems as a major part of their content moderation processes.

A final body of work in the institutional perspective explores the logistics of case-by-case decision-making in platforms' moderation processes. Due to the secrecy surrounding their design, little academic research has been able to study how platforms' proprietary moderation algorithms make decisions; Safiya U. Noble's *Algorithms of Oppression* analyzes algorithms from an institutional perspective, but does not focus specifically on moderation [51].[6] However, extensive research by Sarah Roberts has uncovered the central role of human labor in what she terms "commercial content moderation" [60]. Though platforms are publicly vague about the details of their moderation processes, they have managed until recently to project the image that moderation was handled by algorithms and company employees.[7] Investigative work by Roberts and a handful of journalists has revealed that Facebook, Google, Microsoft, YouTube, and many others now employ or hire as contractors thousands or even tens of thousands of workers from around the world whose job is to click through content that has been flagged (by algorithms and/or humans) to determine whether it is permitted on the platform. Roberts notes that these workers are typically low-status and receive low pay, and are frequently from less developed parts of the world [61, p. 50].

Though all of the above work is foundational in this domain in at least one respect, it is Roberts's work on commercial content moderation that has most directly driven my research. This model for social platform moderation seems unsustainable to me; the structure and (in)visibility of these workers' labor implies that platforms do not

[6] A wide variety of academic research has attempted to develop improved algorithms for detection of problematic behaviors, but this work does not take the platforms' perspective. We discuss this work in the following section.

[7] Gillespie notes that, in its early days, Facebook relied on Harvard students to volunteer their time as moderators. [22, p. 118]

see them as a permanent part of their processes, but rather as a stopgap measure until algorithms improve in quality enough to take over. While these algorithms have become quite effective in identifying spam, fake accounts, and explicit pornography,[8] it is a simple epistemological exercise to show that it is objectively impossible for them to reach full accuracy in identifying fake news or hate speech or cyberbullying. Given this, is our social media future inextricably tied to the labor of armies of commercial content moderators? Or is there another option?

2.3 Systemic problematic behaviors

The second perspective focuses on properties of problematic behaviors online and the users who they involve. Unlike the above perspective, this perspective does not have a unifying focus or clear seminal works. Instead, it can be seen as a loose unification of research into related topics like hate speech, harassment, and cyberbullying.[9] This perspective can be broken into three main (non-exclusive) categories defined by core questions: First, **what are the problematic behaviors that happen online?** Second, **why do these behaviors happen?** And third, **what can be done in response**?

2.3.1 What are the problematic behaviors that happen online?

While more recent work has investigated problematic behaviors as phenomena in and of themselves, the roots of this work come from more ethnographic work on early online spaces. Reid described the impact of the first influx of teenage users on to the internet in the early 1980s [59], focusing on their tendency to push the limits of existing moderation and security technologies to see what they could get away with. In her 1984 book, *The Second Self*, Turkle hypothesized that this behavior results

[8]See accuracy details provided in, e.g., `https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works`

[9]Though I use the term "cyberbullying" here because it is common in the literature, I typically find the term to be counterproductive. It implies a strict distinction between the types of problematic online behaviors exhibited by children and adults, but much research has found that the distinction is frequently not so black and white.

from the obsession of a certain population of (typically male) users with the power they can have over digital systems and other users [80].[10] In parallel with this more ethnographic work, social psychologists were experimenting on the social dynamics of computer-mediated communication. Kiesler, Siegel, and McGuire discussed, among other things, the impact of anonymity on behavior in this medium [35], speculating that a lack of social cues online could lead to "stronger or more uninhibited text" (p. 1125), which, paired with a lack of widespread norms for online communication, might lead to problematic behaviors. Though the primary contribution of this section was identification of important social characteristics of online communication, it has instead been used subsequently to justify the assertion that anonymity directly leads to bad behavior.[11]

Subsequent literature focused on expanded documentation of different types of problematic behaviors online in growing online communities from MUDs (Multi-User Dungeons) to Usenet forums and other online bulletin boards. Donath's work built on Kiesler's work on social signals to explain trolling in Usenet forums as a result of variable ability to prove honesty in communicating information [14]. Sternberg's extensive work on *Misbehavior in Cyber Places* in the 1990s (published as a book in 2012) identified an "infamous triad of troublesome online behavior" of flaming, spamming, and virtual rape [71, p. 77–85]. Though the former two categories had been discussed extensively in early work on problematic behaviors online, the latter category emerged from literature from the mid-to-late 1990s that studied gender dynamics in online communities. This category in Sternberg's analysis builds in large part from a famous incident of online sexual harassment first reported by Dibbell [12] and subsequently discussed in more depth by MacKinnon [42]. Other researchers have identified other similar examples, including Herring et al.'s analysis of trolling in a feminist forum [25].

This "triad" of categories, along with the broad label of "trolling", have remained

[10]Turkle names these users "hard masters" and characterizes them as strong in mathematical and scientific thinking, as opposed to "soft masters" who are creative and artistic.

[11]See `https://web.archive.org/web/20190717010642/https://coralproject.net/blog/the-real-name-fallacy/` for a more detailed discussion.

a focus of social computing literature, though analysis of spam has decreased as spam filters have increased in quality. Kiesler et al. present a variety of examples of these types of behaviors, connecting the 1990s with the 2000s [36]. More recent work includes Phillips's 2015 analysis of Facebook memorial page trolls, which described mobs of users who made use of Facebook features to prey upon the grief of families who had lost loved ones [54].[12] Moor, Heuvelman, and Verleur performed early analysis of flaming in YouTube comments [47], one of the most notorious spaces for classic flaming behaviors on the modern web. Citron, working from a legal perspective, analyzed some of the most harmful sexual harassment (or, per Sternberg, "virtual rape") behaviors online in her work on "revenge porn", the non-consensual distribution of sexual imagery and video of a person in order to shame or harass them [9]. A plethora of other studies have examined the dynamics of these and other types of problematic behaviors, but the above represent notable exemplars of major categories.

2.3.2 Why do these behaviors happen?

Beyond identifying problematic behaviors and their impact, various research has performed studies and experiments to attempt to understand the underlying causes of these behaviors. One of the earliest threads in this work, as mentioned above, comes from Kiesler's work on identity signals in online communication [35] (see also [68] and [70]). The basic premise of this work was that online communication lacked the same contextual cues that face-to-face communication has, from body language to facial expressions to tone and pacing of speech, and that this lack of cues could lead to difficulties in communication. Kiesler, Siegel, and McGuire also noted that the prevalence of relative anonymity online could lead to disinhibition. Donath later presented a 'Social Signaling Theory' based on this general concept, drawing from biological theories on how animals signal to each other combined with an analysis

[12]Notably, Phillips has subsequently argued against the use of the term "troll" because of its lack of specifity and sugar-coating of harassment and other harmful behaviors. See `https://web.archive.org/web/20190620174514/https://kernelmag.dailydot.com/issue-sections/staff-editorials/12898/trolling-stem-tech-sexism/` for more discussion. Phillips and Milner's broader work on nuance of internet culture also provides commentary on the complexities of online communication [55].

of behaviors in Usenet forums [14]. She proposed a distinction between *conventional signals* and *assessment signals*. Conventional signals are easy and cheap to produce but difficult to verify; a person could claim in an online message board to be a professional athlete, but this assertion carries little weight on its own. Assessment signals are costly to produce but easier to verify. On Reddit's r/IAmA board, users hosting a major Q&A must provide concrete evidence to the moderators that they are in fact the person or type of person that they claim to be, and moderators in turn verify this fact to users. The examples of 'trolling' that Donath discusses often result from the inability to verify that users are who they claim to be. Friedman and Resnick note that platforms are often structured in a way so as to make it easy to make new accounts, further exacerbating the problem of signals by making it difficult to track users' history when they seek not to be tracked [16].

In popular discourse, anonymity is often thought of as a cause of bad behavior. In academic writing, Suler's "The Online Disinhibtion Effect" is often cited as justification of this type of assertion [74], but this article is in fact not a causal proof of such an effect but rather a speculative article proposing factors that might contribute to problematic phenomena that are commonly observed online. Literature in experimental psychology has in fact demonstrated that disinhibition is not directly linked to problematic behaviors. The major work in this domain is Reicher, Postmes, and Spears's Social Identity Model of Deindividuation Effects [58], which demonstrates that deindividuation leads to stronger conformity to existing social norms. When deindividuation occurs within an angry mob, these mob members will naturally behave worse than if they were strongly individuated, but deindividuation within a space focused on charity or social support leads to stronger adoption of these positive social norms. For example, much work has shown the prevalence of deeply supportive behaviors in anonymous online cancer support forums [85, 86]. A more recent meta-analysis confirmed this relationship between anonymity and conformity in online contexts [30], finding larger effect sizes in contexts where users were aware of the presence of an outgroup. This latter finding is also predicted by established theory in the social identity paradigm (e.g., [76], [77]).

A more recent body of literature has built on this focus on situationally-present social norms. Álvarez-Benjumea and Winter provide experimental evidence that descriptive norms impact users' behaviors in online comment threads, showing that observing other users' comments being "censored" for containing problematic content led users to post more positive comments [1].[13] While Álvarez-Benjumea and Winter did not find significant impact of injunctive norms (in this case operationalized as rebukes) in their experiment in the context of an online comment thread, Munger finds a significant effect of rebukes on Twitter when the rebuking user is a high-status individual of the same race as the target [48]. These results are both in line with my own past work on imitation, which showed that users imitate each other's positive and negative behaviors, particularly when the originator is in a position of authority, but that observing other users being punished for a particular behavior made them less likely to imitate it [64].[14,15] Cheng et al. found similar results in analyzing the impact of previous comments on users likelihood to engage in problematic behaviors [8] My later qualitative work studying social dynamics in online communities also supports the importance of descriptive norms – moderators find that requiring users to wait a few minutes before participating for the first time allows them to observe what types of behaviors are acceptable, and thus makes them less likely to behave improperly in their first messages [66]. Blackwell et al. demonstrate the complexity of understanding norms for behavior in potentially problematic scenarios by showing that harassment can be perceived as variably justified and deserved depending on the motivation [3]; participants found harassment to be more justified and deserved when the target had supposedly stolen money from an elderly couple, but this effect can be reversed by the intervention of a "bystander" who critiques the harassment.

It is also worth note that diversity in and of itself has been proposed as a contrib-

[13]Kiesler et al., in earlier work, had recommended displaying examples of inappropriate behaviors to clarify norms, which this study operationalizes.[36]

[14]We framed this work as focusing on *imitation* rather than *norms* because of the very short time period over which we analyzed effects, but at the core this effect is about observation of what others are doing.

[15]In this work we also described a "conception effect" loosely based off the work of Wheeler [83], where users become much more likely to engage in a particular behavior once they are reminded of the possibility. To my knowledge this particular effect has not been studied further.

utor to conflict online. In contrast to the classic early internet philosophy that the internet would help people transcend boundaries of race, gender, and nationality,[16] Smith argues that wide cultural diversity and disparate interests, needs, and expectations make conflict more likely and harder to manage [69, p. 160]. I found evidence in my interviews of volunteer community moderators that an increase in diversity of a group's membership (often as a result of growth in membership) often preceded a group's collapse; several Facebook Group moderators told me stories about how political groups they had previously moderated fell apart when they attempted to integrate members of different political orientations. One described a major conflict that occurred when a group she had created dedicated to a niche hobby had grown beyond her urban, liberal, college-aged peers to include an older, more rural, conservative population. Smith concludes with a call to work to find better ways to facilitate more positive interactions within diverse populations online, which, despite having been written more than twenty years ago, is perhaps more relevant today than it has ever been.

The above work all focuses on social factors impacting likelihood to misbehave from social signals to disinhibition (via conformity) to descriptive and injunctive norms. Though my work focuses on these social factors, it is worth mentioning that another thread of research has recently begun to focus on explaining "trolling" via analyzing the personalities of trolls. For example, Lopes found that users with higher psychopathy scores were more likely to support trolling a fake Facebook profile [41], and Stiff found that Machiavellianism and psychopathy predicted likelihood to engage in what he terms 'Facebook surveillance' [72].

2.3.3 What can be done in response?

Research has studied and proposed various responses to these problematic behaviors. Per the previous sections, these responses are largely either social, in that they suggest different ways people can behave, or sociotechnical, in that they suggest new features

[16] See, e.g., the classic cartoon: `https://web.archive.org/web/20190720084744/https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you're_a_dog`

or tools that can shape the way people behave. The former category is often written implicitly or explicitly in previously-discussed work. For example, in my work showing users' imitation of others' behaviors I suggest that moderators have significant power in setting an example for how to behave [64]. Munger's work implies that high-status ingroup users have similar authority in spaces like Twitter [48], and Blackwell et al. note the power of bystanders [3]. In other work, Fox and Tang identified strategies that women already employ in coping with sexual harassment in online games [15], but also note problems associated with each of these strategies; women cannot overcome online harassment simply by coping better. Vitak et al. also study women's strategies for mitigating negative effects of online sexual harassment in a population of undergraduate and graduate women, identifying areas where platforms have been deficient in supporting vulnerable users [81].

The space of research proposing and testing sociotechnical interventions has largely been situated within HCI and CSCW. One direction these interventions have taken is leveraging social support and the help of friends in responding to harassment or other problematic behaviors. Mahar, Zhang, and Karger created a tool called *Squadbox* that makes use of "friendsourced" moderation by allowing users to designate other users to screen their email prior to it arriving in their inbox [43]. Blackwell et al. studied *HeartMob*, a system that allows users to submit examples of harassment that they are facing, after which a pre-established "Mob" of users floods them with supportive comments [4]. My work studying community moderation dynamics also finds support for the efficacy of this type of social support [66]; visible figures in Twitch communities often designate moderators to pre-screen chat messages in a way similar to the email screening in *Squadbox*. Geiger studies a collective approach to mutual moderation via Twitter blocklists, where users work together to coordinate lists of users who they all agree to block [17], a phenomenon further explored in [33].

A popular field of recent study has been the development of algorithms to automatically identify and remove offensive content. Work in this domain has focused largely on spaces like Twitter [5], online news comment sites [50], and other forums [87], though more recent work has looked at platforms including Instagram [40] and

Reddit [7]. These approaches face a number of significant challenges. First, there is no standard way to define problematic content, so these paper typically present a classification schema, a method for detection, and measures for evaluating effectiveness simultaneously.[17] Each paper defines its focus in a slightly different way; Yin et al. detect "harassment", defined as "communication in which a user intentionally annoys one or more others" [87, p. 1]. Nobata et al. focus on hate speech, but do not provide a specific definition by which raters applied this label [50]. Burnap and Williams also focus on hate speech, defining it to raters as content that is "offensive or antagonistic in terms of race ethnicity or religion" [5, p. 227]. Liu et al. detect "hostile" content, defined as "containing harassing, threatening, or offensive language directed toward a specific individual or group" [40, p. 183]. These differing definitions make comparison of the effectiveness of the various approaches virtually impossible. Moreover, as Blackwell et al. note [3, p. 24:3], automated detection systems have been found to be very vulnerable to slight changes in text, where simply changing a single character can completely reverse a comment's sentiment score even in very complex algorithms. Similarly, in studying Instagram's attempts to block access to pro-eating disorder content, Gerrard finds that users have adopted various strategies to actively circumvent Instagram's filters [19].

Relatively little work has quantitatively analyzed the impact of platform moderation decisions. Chandrasekharan et al.'s work studying the impact of Reddit's decision to ban certain forums is noteworthy in this regard. This work found that this exclusion led to a decrease in behaviors previously characteristic of the excluded communities [6], though it did not analyze changes in problematic behaviors that were not directly associated with these communities.

Broadly, these approaches largely presume a baseline level of problematic behaviors that can be addressed through removal or social response. Very little literature has looked at sociotechnical interventions designed to proactively reduce the preva-

[17] Chandrasekharan et al. avoid the problem of definition by using external communities' definitions of problematic behavior to define rules for a new community. This results in a classifier that, as expected, works not from consciously defined rules but from the aggregation of prior user moderation decisions [7].

lence of these behaviors. I have begun to explore potential interventions through analysis of the impact of chat moderation modes [64], which prohibit certain types of content from being posted, and affective priming via the use of specially-designed CAPTCHAs [62], but I discuss this direction in more depth in my proposed work.

2.4 Communities' self-moderation

Since the beginning of the internet, a large portion of social interactions online have been structured within online communities. The first form of these was the Bulletin Board System (BBS), as described in Reid [59], which were created and run primarily by academics and researchers. Usenet forums, similar in structure to BBS but stored in a more decentralized fashion across many servers, emerged in the early 1980s (see e.g., [14] and [73]), accommodating a much broader and more casual user-base. Multi-User Dungeons (later called Multi-User Domains or Multi-User Dimensions, abbreviated as MUDs or MU*), which were game-based online social environments often built around a role-playing theme, emerged in the 1980s but grew significantly in popularity during the 1990s [12, 59, 69, 71]. Though the 2000s saw rapid growth in network-style social media like Myspace and Facebook (followed by Instagram, Twitter, and others in the 2010s), the decade also saw the introduction of virtual communities like Wikipedia [18, 31] and Reddit [45]. A variety of smaller community-based platforms have emerged in recent years, including Twitch [79], Discord (Kiene and Jiang forthcoming), and an increased focus from Facebook on its Groups feature.[18]

These platforms have fundamentally different characteristics than the types of platforms that Gillespie, Klonick, and others have highlighted. While platforms like Twitter, Instagram, YouTube, and Facebook (excepting Facebook Groups) are governed almost exclusively by company policies and their implementations, community-centric platforms like Twitch, Wikipedia, and Reddit, are governed almost exclusively by users. This type of governance usually manifests in the form of groups of volun-

[18]Though most of these platforms have been extensively studied from a variety of perspectives, I have selected work in this paragraph that is primarily ethnographic in order to provide starting points for each platform. Work from different perspectives is discussed later in this section.

teer user moderators, users who either create communities on these platforms or who are chosen by these founders. These volunteer moderators have access to baseline tools on the platform to ban or time out users who misbehave and to remove offending content, and each platform provides additional tools that are variants on similar themes. The importance of volunteer moderators' work has been clearly established in the literature; many users prefer to participate in spaces that are well-moderated [84], and this type of moderation can increase the quality of users' contributions [10] and help steer communities through periods of turbulence [34]. Though platform administrators do technically have "veto power" over volunteer moderators' decisions in that they can choose to remove users, content, or entire communities without input from these volunteers[19], the relationship between platforms and volunteer moderators is typically distant if not nonexistent. The vast majority of volunteer moderators on these platforms never encounter interference from platform administrators in the way they govern their communities [66]. In this section I expand on the above themes, exploring the various frameworks that have been developed to explain community moderation processes online.

In a wide-ranging synthesis of research on online communities published in 2012, Kraut and Resnick identify five major challenges that community leaders face: (1) Encouraging Contribution; (2) Encouraging Commitment; (3) Regulating Behavior; (4) Dealing with Newcomers; and (5) Starting New Communities [38]. Though what Smith describes in her work on theories of conflict management as an "integrationist" perspective on conflict would see the moderators' role as situated primarily in the third of these five categories, the "radical" perspective on conflict that Smith highlights understands moderation as part of a broader process of community governance [69, p. 135]. Though the term "moderator" evokes imagery of bans and removals, volunteer community moderators are actually more akin to community leaders. They are responsible for building communities, often from their inception, and guiding them toward positive cultures of social interaction. Of these five categories, two have received the most focus in studies of the practices of volunteer community moderation:

[19]With the obvious exception of Wikipedia

dealing with newcomers and *regulating behavior*.

2.4.1 Dealing with Newcomers

In her studies of the MU* MicroMUSE performed in the early-to-mid 1990s, Smith identified a nuanced and evolving process for integrating newcomers into the community [69, p. 148]. Though initially more lax and open, this process changed significantly several years into MicroMUSE's operation in response to conflict that followed a rapid growth in the community's membership. Server administrators significantly restricted the commands that visitors could use to interact with other users, and required that all new members to receive "sponsorship" from two existing members after a period of socialization. A program for "mentorship" of these newcomers was also created, providing liasons between newcomers and the main user-base. All of these processes appear in similar forms in modern platforms. Facebook Groups are often set by their moderators to be "closed" or "secret", with the former requiring users to request to join or be invited, and the latter only visible to users who are specifically invited.[20] "Followers-only mode" on Twitch requires users to have been present for a certain amount of time before they are allowed to post [66], and Automoderator settings on Reddit can also prohibit new users from posting in certain communities [32]. Wikipedia maintains an "Adopt-a-user" mentoring program that pairs new users with more experienced Wikipedians [49].

2.4.2 Sociotechnical approaches for moderating behavior

The mechanisms that moderators use for regulating behavior of established (and new) members have also been studied at length, with distinctions typically made between social and sociotechnical approaches as discussed in the previous section. The latter category is often more visible – volunteer moderators frequently make use of bans and content removals, and it is unlikely that a user could participate in groups online for any significant length of time without seeing these tools used [66]. Tools for

[20] `https://web.archive.org/web/20190617064702/https://www.eff.org/deeplinks/2017/06/understanding-public-closed-and-secret-facebook-groups`

automated removal are also common, though their detection algorithms are typically simple and rules-based [32, 63]. Crawford and Gillespie's aforementioned work on the use of flags to highlight problematic behaviors focused on their use by platforms, flagging and reporting tools on platforms like Reddit can be used to send reports to communities' moderators rather than platforms, and moderators in busy spaces often rely significantly on user reports to focus their attention [66].[21]

2.4.3 Social strategies for encouraging more positive behaviors

Though much focus has been placed on the tools that moderators use to manage their communities, the vast majority of the 56 moderators I interviewed from Twitch, Reddit, and Facebook Groups found social approaches to be more important or central to their moderation practices [66]. This was particularly the case in smaller communities where relationships and social structures could be established and maintained. These strategies are implicit in much prior work on moderation (e.g., [69, 71]), but my research was the first to formally place these strategies within a broader framework of practices in moderation. I detail these results more in the following chapter, but I provide two examples here: first, many moderators were conscious of the importance of setting an example. They realized that their status made them particularly visible within the group and that other users observed their behavior, whether consciously or not, as a signal of what types of conduct are appropriate. This matches findings in [64], which showed the influence of moderators in shaping regular users' future behaviors. Second, moderators used an escalating set of responses to problematic behaviors which began as social responses and escalated into technical responses (i.e., time-outs or bans). Moderators would often first let an offender know that their conduct was inappropriate via a private message or brief comment response. On Reddit and Facebook, this could be intertwined with a technical response, as moderators can elect not to approve top-level posts if they don't meet certain requirements. If

[21] These tools are all built around sanctions, and it is worth noting that there are very few features on any of these platforms that allow moderators to provide positive attention or rewards to users who behave well.

an offender continued to behave poorly, moderators would either issue a stern, direct warning or a brief time-out. This would eventually be followed by a ban from the space. Depending on the severity of the infraction or the nature of the offender, moderators sometimes skipped steps in this process. For example, if an offender was perceived to be a bot rather than a human user, moderators would typically skip directly to banning it; they found it unlikely that a bot would reform its conduct. Similarly, users who displayed extreme behaviors (e.g., aggressive racial slurs, rape threats) would often be immediately banned.

2.4.4 Strategies and structures for governance in volunteer moderation

One of the most important parts of the job of a volunteer community moderator is to decide what conduct is appropriate and what conduct is not. This process mirrors the processes that occur as platforms develop, per Gillespie's descriptions in [22], except typically on a much smaller scale. Work on the development of rules or codes of conduct in online communities has made it clear that the clear display of rules is important and effective in reducing misbehavior [36, 46]. Sternberg, drawing from literature both in CSCW and the sociology of deviance, identifies three rule-related social processes in online communities: *rule-breaking*, *rule-making*, and *rule-enforcement* [71, p. 155-169]. She notes that these processes take place in variable orderings; a simplistic assumption would be that rules are created for a space, and when users break or threaten to break these rules, moderators enforce them. However, in reality rules are often created after a perceived offense has already occurred. Moderators simply do not have the foresight to anticipate all of the different ways in which users might behave that would prove harmful to the community. In one of the interviews I performed, the moderator described having to create a rule to forbid users from posting advertisements for t-shirt sales, which had become popular on Facebook due to the ubiquity of a particular ad that purported to sell t-shirts with text customized

based on the user's data traces.[22] Another moderator I interviewed that manages a subreddit dedicated to a popular hobby described having to ban pictures of product boxes. While he understood that users were excited to show off their purchases, the moderators felt that the flood of posts that contained only a picture of a product in its box were reducing the overall quality of content in the community. The philosophies moderators draw on for rule-writing are often a combination of personal values and prior experience within online spaces. While one of the fundamental philosophical conflicts of the internet has been between the cyber-libertarian perspective advocating for unrestricted speech and self-governance and others who seek a more civil online society,[23] rule-writing often comes down to a combination of pragmatism and conversations between moderators about what type of space they want to create.

The above examples show how rules can be "broken" before they are written, but, as Sternberg notes, in some cases rules are even enforced before they are written. I present examples from MacKinnon [42] and Smith [69] in the following paragraphs where moderators took action against what they perceived to be an egregious offense before the rules related to such offenses were established, leading to extensive community debates. In my work I found that this was often an issue of either a lack of specificity in rules or a conflict between moderators' values and the wording of existing rules. Twitch moderators stated (almost always verbatim) that the fundamental rule across all of Twitch is "don't be a dick" [66]. Many recounted instances where they removed users for "being a dick", even though it was clear that each moderator had a slightly different interpretation of what this meant. In some cases this type of vagueness led to rewriting or clarification of rules, but moderators in Facebook Groups and on Reddit were more likely than Twitch moderators to focus specifically on iterating wording of rules. This may be a result of the slower (asynchronous) pace of conversation on these platforms; while on Twitch moderation actions are often only

[22] `https://thenextweb.com/socialmedia/2019/04/08/subreddit-targeted-facebook-shirt-ads/`

[23] John Perry Barlow's "A Declaration of the Independence of Cyberspace" is a foundational document in the cyber-libertarian perspective from the late 1990s (`https://web.archive.org/web/20190721002225/https://www.eff.org/cyberspace-independence`), and Marwick's *Status Update: Celebrity, Publicity, and Branding in the Social Media Age* traces this philosophy through the 2000s and early 2010s [44]. See also [20] for a modern quantitative analysis of this philosophical clash in action

visible for seconds because the flow of the chatroom moves them off-screen, there is often plenty of time for users to debate and argue about details of moderation actions.

As is the case with platforms' governance practices, volunteer moderators' decisions are typically opaque and do not often involve community input [66]. None of the major modern platforms are built with any tools for making democratic decisions in moderation,[24] and moderators I interviewed were pessimistic about the possibility for any sensible decisions to come from community input. In several cases, these moderators had tried to be transparent about these processes and gathered community input, but none of these moderators had maintained this transparency in the long term. One reported that they found it significantly easier to moderate non-transparently because users don't complain about what they can't see or don't know about.

Though these oligarchic, dictatorial, or feudal approaches to governance have been the default at least since the rise of Reddit, early online communities experimented with various models for more democratic governance. MacKinnon [42] describes an incident originally reported by Dibbel [12], where one user of a fantasy-themed MUD called "LambdaMOO" sexually harassed other users by using a "voodoo-doll" subprogram to make it appear as if their characters were performing extreme sexual acts. The responses to this behavior were technical and administered by power users – first a time-out, fantasy-themed as a magical cage, and then a ban via permanent transformation into a "toad". However, this eventual ban came following an extensive discussion at a large gathering of community members, though the final decision was made alone by a moderator judging the consensus reached. Following this incident, the community's "archwizard" (who would now be referred to as a "head moderator") developed a system for petitions and ballots so that important issues raised by any user could be put to popular vote if the interest was great enough:

> *And though some anarchists grumbled about the irony of [the archwizard]'s dictatorially imposing universal suffrage on an unconsulted populace, in*

[24] Wikipedia's decision-making processes could in some senses be called Democratic, but in reality they are dominated socially and technically by a very small fraction of users. The case of governance on Wikipedia has been analyzed at length in [24] and [67]

general the citizens of LambdaMOO seemed to find it hard to fault a system more purely democratic than any that could ever exist in real life. Eight months and a dozen ballot measures later, widespread participation in the new regime has produced a small arsenal of mechanisms for dealing with the types of violence that called the system into being.[12]

Though this LambdaMOO incident is by far the most famous and most cited, it was not an isolated case. Smith describes another incident in a MU* called "MicroMUSE", a community of a few thousand users focused around education with a futuristic theme, where a teenage user named Swagger built an "Orgasm Room" filled with sex objects where he brought female players [69, p. 139-141]. Upon discovering this, a moderator immediately "nuked" Swagger's character, completely deleting it from the database along with all of his belongings. While arbitrary decisions to ban a user happen regularly now across numerous platforms without any notable response from other users, the residents of MicroMUSE revolted in defense of Swagger's perceived right to an opportunity to defend himself before being removed, and two staff helped Swagger re-create his character. This revolt led to the organization of a community-wide town hall to discuss decision-making processes on the MU*. Citizens called for various reforms including elections for moderators, the establishment of a justice system, and checks on the power of moderators. A month later, the MU* adopted a new governing charter that created a "Citizens Council", established procedures for handling incidents of misbehavior, and provided a limited right of appeal. Swagger's case had a less democratic ending – prior to establishment of this new charter, a vote by appointed moderators affirmed the original decision to "nuke", and Swagger's reconstructed character was nuked again and he was banned for two months. Following a successful petition for readmission after two months, he was eventually permanently banned for additional behavioral violations. A similar fate befell a new MUSE established as a community to protest the actions of the MicroMUSE moderators, which was found to violate the policies of its system administrator's university and closed. Though these end results were reached via less-than-democratic means, Smith describes several other examples of the effective-

ness of the new, more democratic processes in responding to problematic behaviors [69, p. 148-158].

The historical processes through which these more democratic approaches to moderation have been lost have not been fully explored, but the transition from (relatively) independently-operated online communities to the establishment of major social media conglomerates is likely a factor. The aforementioned MUDs were comparatively small communities with at most a few thousand members. While members were often fairly technically-savvy, these communities had at most a handful of developers who might choose to work to create tools and features in their spare time. Because of this, communities were often left to handle disputes using social rather than technical means. Though major modern platforms have been slow to create novel features for moderation, these features are far more integrated into platform interfaces and user workflows. It is also possible that the near-infinite options for communities to join that modern platforms have provided has led to a decrease in users who are strongly committed to any single Facebook Group, subreddit, or other community, though I know of no work that has attempted to answer this question longitudinally. A final possibility, supported by the evidence Marwick provides in mapping the philosophies of employees of social media companies [44], is that the designers and developers who have created these platforms are philosophically inclined to believe in the value of technical solutions to social problems, which could have led to framing moderation as problem best addressed by the development of more sophisticated technical tools rather than by facilitating more sophisticated social processes.

2.5 Approaching volunteer community moderation through social identity

Research in Computer-Supported Cooperative Work (CSCW) has explored the evolution of collaboration and cooperation in online systems since the early years of social

computing.[25] Findings have guided the development of systems and the formulation of theoretical models explaining the ways humans engage with them. The social identity perspective (SIP)[26] explains how people organize themselves into and within groups and how they treat both members of their own groups and members of other groups. It identifies factors that cause variation in levels of attachment to and identification with a group, predictors of intra-group and intergroup conflict and the approaches groups take in responding to conflict, and variables that explain how social structures emerge within groups. Because of its focus on the major social dynamics of groups, I have used it to frame and motivate much of my work in moderation in online communities. In this section I briefly review the four principles of social identity that are most applicable to this work and discuss how they might be used.

Principle 1: The identities we tend to embody are those that are the most accessible and have the best "fit" within a given situation.

According to this view, the distinction between personal and social identity reflects the aspects of the self that arise when one makes interpersonal (me versus not me) versus intergroup (us versus them) comparisons and judgments. In addition, this view posits that identity itself is context-dependent, with an inverse relationship between the salience of one level of identity versus the other [52]. Indeed, the principles of accessibility and fit, as elaborated in [28], describe the types of identities that are likely to be most salient at any given time. People draw on accessible identities - those that are important to the individual and connected to their self-concept, and those that are activated by current goals or social context (e.g., the composition of one's immediate context). Race and gender are common identity categories that match both of these criteria, particularly for oppressed or marginalized groups, because of

[25]This section draws significantly from Seering et al. 2018, "Applications of Social Identity Theory to Research and Design in Computer-Supported Cooperative Work" [65]

[26]The term "SIP" includes the original social identity theory first proposed by Tajfel and Turner in the 1970s, but also broadly refers to theories built on this framework, including self-categorization theory, the social identity theory of leadership, and the social identity model of deindividuation effects (SIDE) and their various subprinciples. We refer to all of these under the general social identity umbrella.

their likelihood to be important to the individual and because of how frequently they arise in everyday life.

Principle 2: Individuals consistently favor groups and identities with which they affiliate over competing or contrasting groups.

The minimal group paradigm [13], one of the earliest areas of Social Identity Theory, explores the smallest possible conditions that are required to cause intergroup differentiation and ingroup favoritism. In one famous experiment [2], participants were assigned to groups randomly, based on the result of a coin flip, and told to allocate points between their own group and the other group. Participants did not know who else within the session was in their group; the groups had no history and no future, and were based on meaningless criteria; no particular competition was suggested between the groups; and points were meaningless and carried no inherent value. Nonetheless, participants still consistently allocated more points to their own group in all variations of the experiment. Even in groups formed on a random or arbitrary basis, patterns of ingroup favoritism consistently emerged. When the strength of affiliation between group members increases, ingroup favoritism only increases; as work on Social Identity has firmly established, people like others who are similar to them in salient ways and are innately inclined to form group boundaries when contexts make social categories salient.

The original core premises of social identity theory were derived from Tajfel and Turner's work on intergroup relations [78]. Though the Social Identity Perspective no longer relies on the concept of minimal groups, a wide variety of studies have shown similar effects across different domains, particularly in the context of stereotypes. Several conditions must be met for these processes to take place [78]. Individuals must have internalized membership in the relevant group as part of their self-concept. They also must have cause for comparison, and comparison must occur across attributes that matter in a given context (e.g., gender in a science classroom, political affiliation in a Facebook group centered on immigration issues) and with an outgroup that is relevant to the comparison and situation. Recent research suggests that the attributes

that hold the most weight in comparisons are ones pertaining to perceived morality (rather than perceived competence or sociability) [39].

Principle 3: Anonymity leads to behaviors more strongly prototypical of group norms.

As previously discussed, the Social Identity Model of Deindividuation (SIDE) was developed starting in the 1990s as a counterpoint to earlier work on the psychology of mobs [58]. Reicher, Spears, and Postmes traced the concept of deindividuation in psychology back to the work of LeBon on crowd psychology, published in France in 1895. The core of LeBon's theory was that to be in a crowd was to lose one's individuality and thus any sense of individual responsibility, and therefore to succumb to base behaviors. Further well-known work on anonymity and deindividuation such as the Stanford Prison Experiment, reinforced this widely held view of deindividuation [88].

The SIDE model emerged as a critique of these interpretations, and was founded on a number of experiments that showed that deindividuation (specifically via anonymity in face-to-face conditions) led experimental participants to behave in ways that were more in line with norms for their group, regardless of whether those norms were pro- or anti-social [58]. These findings have been extended to CMC contexts [56, 57]. Thus, in cases where norms are founded on harassment or disruptive behaviors, anonymity can lead to extreme negative behaviors, while in supportive online communities, anonymity promotes pro-social outcomes, such as greater compassion and empathy.

Principle 4: In groups, the leaders who emerge are the members who are most prototypical of the group's norms

In contrast to theories of leadership that suggest that leaders are those who have personality traits relevant to leadership or that group members with the greatest access to resources, the social identity theory of leadership argues that group members who are the most prototypical of group norms emerge as leaders [29, 26]. This

process has three defining phases: first, self-categorization creates a spectrum of prototypicality within the group, with certain members deemed to be more prototypical than others. Second, per the social attraction hypothesis, more prototypical group members are liked more than less prototypical members, and are thus able to exercise influence over other group members because individuals are more likely to help and support people that they like [27]. As the group reaches general (though often not consciously discussed) consensus regarding who is most liked, this person becomes more and more able to exercise power in ways that cement their status. Third, group members make an attribution error [78] by overattributing a leader's position to their personality characteristics rather than their prototypicality, reinforcing the belief that the leader possesses a particular disposition that helped them achieve their status within the group. It is important to note that, while individuals' cognitive representations of prototypical qualities of groups and group norms are conceptually very similar, they are not entirely the same- group norms are better conceptualized as the the aggregation of individual prototypes into collectively agreed upon group prototypes [28].

The literature in the first part of this chapter represents a broad cross-section of literature in the moderation research space, and in this last section I have provided a complementary perspective drawn from social identity theory. It is within the community-driven moderation space that I aim to situate my dissertation.

Bibliography

[1] Amalia Álvarez-Benjumea and Fabian Winter. Normative Change and Culture of Hate: An Experiment in Online Environments. *European Sociological Review*, 2018.

[2] Michael Billig and Henri Tajfel. Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, 3(1):27–52, 1973.

[3] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. When online harassment is perceived as justified. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[4] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACM on Human-Computer Interaction*, 1(CSCW), 2017.

[5] Pete Burnap and Matthew L Williams. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy and Internet*, 7(2), 2015.

[6] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *PACM on Human-Computer Interaction*, 1(CSCW), 2017.

[7] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The Bag of Communities Approach: Identifying Abusive Behavior Online with Preexisting Internet Data. *Proceedings of the 2017 SIGCHI Conference on Human Factors in Computing Systems*, pages 3175–3187, 2017.

[8] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017.

[9] Danielle Keats Citron. *Hate Crimes in Cyberspace*. Harvard University Press, 2014.

[10] Dan Cosley, Dan Frankowski, Sara Kiesler, Loren Terveen, and John Riedl. How oversight improves member-maintained communities. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05*, pages 11–20, 2005.

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

[12] Julian Dibbell. A rape in cyberspace: How an Evil Clown, a Haitian Trickster Spirit, Two Wizards, and a Cast of Dozens Turned a Database Into a Society. *The Village Voice*, pages 1–7, 1993.

[13] Michael Diehl. The minimal group paradigm: Theoretical explanations and empirical findings. *European review of social psychology*, 1(1):263–292, 1990.

[14] Judith Donath. Identity and Deception in the Virtual Community. In Marc A Smith and Peter Kollock, editors, *Communities in Cyberspace*, pages 27–58. Routledge, London and New York, 1st edition, 1999.

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

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

[17] Stuart Geiger. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6), 2016.

[18] D. Geiger, R. S., & Ribes. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work. ACM*, 2010.

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

[20] Anna Gibson. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society*, 5(1):205630511983258, jan 2019.

[21] Tarleton Gillespie. The Politics of 'Platforms'. *New Media & Society*, pages 1–3, 2010.

[22] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.

[23] James Grimmelmann. The Virtues of Moderation. *Yale Journal of Law and Technology*, 17(1), 2015.

[24] Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688, 2013.

[25] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society*, 18(5):371–384, 2002.

[26] Michael A Hogg. A social identity theory of leadership. *Personality and social psychology review*, 5(3):184–200, 2001.

[27] Michael A Hogg and Elizabeth A Hardie. Social attraction, personal attraction, and self-categorization-, a field study. *Personality and Social Psychology Bulletin*, 17(2):175–180, 1991.

[28] Michael A Hogg and Scott A Reid. Social identity, self-categorization, and the communication of group norms. *Communication theory*, 16(1):7–30, 2006.

[29] Michael A Hogg and Deborah I Terry. Social identity and self-categorization processes in organizational contexts. *Academy of management review*, 25(1):121–140, 2000.

[30] Guanxiong Huang and Kang Li. The effect of anonymity on conformity to group norms in online contexts: A meta-analysis. *International journal of communication*, 10:18, 2016.

[31] Dariusz Jemielniak. *Common knowledge?: An ethnography of Wikipedia*. Stanford University Press, 2014.

[32] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2019.

[33] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):12, 2018.

[34] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. Surviving an eternal September: How an online community managed a surge of newcomers. In *CHI '16*. ACM, 2016.

[35] S Kiesler, J Siegel, and T W McGuire. Social psychological aspects of computer-mediated communication. *American Psychologist, 39(10)*, pages 1123–1134, 1984.

[36] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design*, 2012.

[37] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 2018.

[38] Robert E Kraut and Paul Resnick. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press, 2012.

[39] Colin Wayne Leach, Naomi Ellemers, and Manuela Barreto. Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, 93(2):234, 2007.

[40] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. Forecasting the presence and intensity of hostility on Instagram using linguistic and social

features. In *International AAAI Conference on Web and Social Media*, volume 91, pages 181–190, 2018.

[41] Barbara Lopes and Hui Yu. Who do you troll and Why: An investigation into the relationship between the Dark Triad Personalities and online trolling behaviours towards popular and less popular Facebook profiles. *Computers in Human Behavior*, 77:69–76, 2017.

[42] Richard MacKinnon. Virtual rape. *Journal of Computer-Mediated Communication*, 2(4):0–0, 1997.

[43] Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *CHI '18*, Montreal, QC, Canada, 2018. ACM.

[44] Alice E Marwick. *Status update: Celebrity, publicity, and branding in the social media age*. Yale University Press, 2013.

[45] Adrienne Massanari. *Participatory Culture, Community, and Play: Learning from Reddit*. Peter Lang, 2015.

[46] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 2019.

[47] Peter J. Moor, Ard Heuvelman, and Ria Verleur. Flaming on YouTube. *Computers in Human Behavior*, 26(6):1536–1546, 2010.

[48] Kevin Munger. Tweetment Effects on the Tweeted : Experimentally Reducing Racist Harassment. *Political Behavior*, 2016.

[49] David R Musicant, Yuqing Ren, James A Johnson, and John Riedl. Mentoring in wikipedia: a clash of cultures. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 173–182. ACM, 2011.

[50] Chikashi Nobata and Joel Tetreault. Abusive Language Detection in Online User Content. In *WWW 2016*, pages 145–153, Montreal, QC, Canada, 2016. ACM.

[51] Safiya Umoja Noble. *Algorithms of Oppression*. NYU Press, 2018.

[52] Rina S Onorato and John C Turner. Fluidity in the self-concept: the shift from personal to social identity. *European Journal of Social Psychology*, 34(3):257–278, 2004.

[53] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Group '16*, pages 369–374, Sanibel Island, FL, USA, 2016. ACM.

[54] Whitney Phillips. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press, 2015.

[55] Whitney Phillips and Ryan M Milner. *The ambivalent Internet: Mischief, oddity, and antagonism online*. John Wiley & Sons, 2018.

[56] Tom Postmes, Russell Spears, and Martin Lea. Breaching or building social boundaries? side-effects of computer-mediated communication. *Communication research*, 25(6):689–715, 1998.

[57] Tom Postmes, Russell Spears, Khaled Sakhel, and Daphne De Groot. Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, 27(10):1243–1254, 2001.

[58] S. D. Reicher, R. Spears, and T. Postmes. A Social Identity Model of Deindividuation Phenomena. *European Review of Social Psychology*, 6(1):161–198, 1995.

[59] Elizabeth Reid. Hierarchy and Power: Social Control in Cyberspace. In Marc A Smith and P Kollock, editors, *Communities in Cyberspace*, pages 107–134. Routledge, London and New York, 1st edition, 1999.

[60] Sarah T Roberts. Commercial Content Moderation: Digital Laborers' Dirty Work. In S U Noble and Tynes B., editors, *The Intersectional Internet: Race,*

Sex, Class and Culture Online, pages 147–160. Peter Lang Digital Formations series, 2016.

[61] Sarah T. Roberts. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday*, 2018.

[62] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–14, Glasgow, Scotland, UK, 2019. ACM.

[63] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The social roles of bots: Evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):157, 2018.

[64] Joseph Seering, Robert E Kraut, and Laura Dabbish. Shaping Pro and Anti - Social Behavior on Twitch Through Moderation and Example - Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017.

[65] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. Applications of social identity theory to research and design in computer-supported cooperative work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):201, 2018.

[66] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society*, 2019.

[67] Aaron Shaw and Benjamin M Hill. Laboratories of oligarchy? how the iron law extends to peer production. *Journal of Communication*, 64(2):215–238, 2014.

[68] Jane Siegel, Vitaly Dubrovsky, Sara Kiesler, and Timothy W. McGuire. Group processes in computer-mediated communication. *Organizational Behavior and Human Decision Processes*, 37(2):157–187, 1986.

[69] Anna DuVal Smith. Problems of Conflict Management in Virtual Communities. In Marc A Smith and P Kollock, editors, *Communities in Cyberspace*, pages 135–166. Routledge, London and New York, 1st edition, 1999.

[70] Lee Sproull and Sara Kiesler. Reducing social context cues: Electronic mail in organizational communication. *Management science*, 32(11):1492–1512, 1986.

[71] Janet Sternberg. *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet*. Rowman & Littlefield, 2012.

[72] Chris Stiff. The dark triad and facebook surveillance: How machiavellianism, psychopathy, but not narcissism predict using facebook to spy on others. *Computers in Human Behavior*, 94:62–69, 2019.

[73] Allucquère Rosanne Stone. Will the Real Body Please Stand Up? In Michael Benedikt, editor, *Cyberspace: First Steps*, pages 81–118. MIT Press, 1991.

[74] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[75] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication*, 13(18), 2019.

[76] H. Tajfel and J. C. Turner. The social identity theory of intergroup behavior. In J. T. Jost and J. Sidanius, editors, *Key readings in social psychology*, pages 276–293. Psychology Press, 1986.

[77] Henri Tajfel and John C Turner. An integrative theory of intergroup conflict. In William G Austin and Stephen Worstel, editors, *The social psychology of intergroup relations*, pages 33—-47. Brooks/Cole Pub. Co, 1979.

[78] Henri Tajfel and John C Turner. An integrative theory of intergroup conflict. *The social psychology of intergroup relations*, 33(47):74, 1979.

[79] T.L. Taylor. *Watch Me Play: Twitch and the Rise of Game Live Streaming*. Princeton University Press, 2018.

[80] Sherry Turkle. *The Second Self*. MIT Press, Cambridge, MA, USA, 1st edition, 1984.

[81] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1231–1245. ACM, 2017.

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

[83] Ladd Wheeler. Toward a theory of behavioral contagion. *Psychological Review*, 73(2):179, 1966.

[84] Kevin Wise, Brian Hamman, and Kjerstin Thorson. Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication*, 12(1):24–41, 2006.

[85] Diyi Yang, Zheng Yao, and Robert Kraut. Self-disclosure and channel difference in online health support groups. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[86] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 31. ACM, 2019.

[87] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7, 2009.

[88] Philip G Zimbardo. On the ethics of intervention in human psychological research: With special reference to the stanford prison experiment. *Cognition*, 1973.

Chapter 3

Moderator Engagement and Community Development in the Age of Algorithms

3.1 Introduction

Recent attention to moderation in both research and public discourse has focused on company-driven removal of unwanted content at scale and the corresponding responsibilities of platforms [15, 8].[1] For example, Roberts and Gillespie [24, 8] identify and discuss the use of many thousands of commercial content moderators whose job is to filter through an endless stream of content and remove what is deemed unacceptable on a given platform. While the politics and mechanics of content detection and removal strategies have taken center stage, the labor done by users to moderate their own communities has not received as much attention. Sites like Wikipedia, Reddit, Twitch [17], and Facebook Groups rely on their own users to do the vast majority of moderation work from the bottom-up, creating a significantly different dynamic than in spaces where moderation is driven top-down by company policy. User-driven moderation is an intensely social process that is core to community development.

[1] This chapter is based primarily on my paper, "Moderator Engagement and Community Development in the Age of Algorithms" [26].

The recent emphasis on scale follows a divergence in the research landscape on moderation as a whole. While early research focused on misbehavior as a group-level phenomenon in user-organized and user-managed online spaces such as Usenet newsgroups, Internet Relay Chat (IRC), and Multi-User Domains (MUDs) [28], more recent work has addressed misbehavior as a platform-level phenomenon following the rise of centrally-organized and managed online spaces like Facebook and Twitter [1, 15, 8]. Despite the shifting focus, many self-governing online communities continue to thrive and, in some cases, are expanding rapidly. In this work, we detail a comprehensive model of how volunteer human moderators govern their communities in the age of algorithms.

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

We identify and describe three interconnected processes that drive moderation and governance in these communities. First, we describe the processes of becoming a moderator, including appointment into the role and development over time. Second, we detail the processes for handling incidents that arise both proactively and reactively. Finally, we describe the decision-making processes for modifying how communities are run. We focus on these three main themes because volunteer moderation is responsible for shaping the online experiences of billions of users distributed among many millions of communities. Past research has explored pieces of these processes, usually on one specific platform. This systematic description of the full process of governance across diverse platforms can inform future research on the effectiveness of different strategies, guiding development of better tools that support, rather than

supplant, the judgment of users.

3.2 Twitch, Reddit, and Facebook

The social dynamics of online communities have been explored from numerous angles. In this work, we focus on informal public spaces where people meet to share interests, converse, and build a community [22], emphasizing the social nature of the spaces and the emergent community-building processes. Facebook Groups, Reddit subreddits, and Twitch channels [10] all meet these criteria, while network structures such as Twitter followers or Facebook friends currently do not. Though there are many online communities that match these descriptions, we chose Twitch, Reddit, and Facebook as three of the largest community-based sites with significantly different features and cultures. Facebook reports over 2 billion monthly active users[2], and Reddit[3] and Twitch[4] report hundreds of millions. While Twitch is much younger than Facebook and Reddit, it has been the subject of much research on community dynamics in the past five years [10, 29, 12].

Each of these platforms hosts different types of communities with different feature-structures. Reddit communities take the form of text-based discussion forums, where visibility of content is determined by voting [19]. Twitch communities are chatrooms built around interaction with a single specific user, the "streamer," who appears on a live video stream [10]. All three platforms provide basic algorithmic tools for handling common misbehaviors. Reddit offers AutoModerator, a bot that proactively catches messages based on user-chosen settings for moderators to review later. Twitch's AutoMod functions similarly, though it requires more immediate attention due to the synchronicity of conversation on Twitch. Facebook's most commonly used algorithmic tool is its automatic flagging of group join requests from suspected spam accounts. However, despite the presence of these tools, the vast majority of moderation decisions

[2] *Facebook's grim forecast: privacy push will erode profits for years.* Reuters. Retrieved from https://www.reuters.com/article/us-facebook-results/facebook-misses-estimates-on-monthly-active-users-idUSKBN1KF2U5

[3] *The conversation starts here.* Retrieved from https://www.redditinc.com/

[4] *Audience.* Retrieved from http://twitchadvertising.tv/audience/

are still made by users either manually or through independently developed bots.

In contrast to both Twitch and Reddit, which are platforms built to host communities, Facebook hosts Groups as a complement to the site's primary social network function. Facebook scales its content moderation by designing algorithms and employing thousands of commercial content moderators to tackle its massive network. [24, 8, p. 120-124].

3.3 Moderation and meaningful communities

Much work in the study of moderation has focused on the specific problems that occur and how they are handled, such as the vexing question of how to deal with "trolls" that plague an online community [11]. Early work analyzed misbehaviors that appeared in communities that were dominantly user-run, such as Usenet newsgroups, MUDs and, more recently, Wikipedia. In a review of online misbehavior in the early social web, Sternberg identifies an "Infamous Triad" of flaming, spamming, and virtual rape [28, p. 77-85]. Spam and broad incivility remain problematic on all types of platforms, and more recent work on online harassment has explored targeted attacks often focused against particular identity groups [4].

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

fill.

Though user and group-level strategies for moderation continue to be studied [14, 25, 13], much recent work has begun to study misbehavior as a network-level phenomenon that can be dealt with using a top-down approach. This represents an implicit shift from managing misbehavior to filtering content. The increasing application of machine learning methods to social computing problems offers a solution for detecting a wide variety of negative behaviors at scale. [8, p. 52-63] Albeit imperfect and often difficult to define, these approaches are important as platforms continue to seek scalable moderation methods. To supplement these algorithms, major platforms also hire thousands of commercial content moderators to sort through suspect content [24, 8]. These top-down approaches emphasizing commercial content moderation lead to several additional questions: How do user-moderators make use of or interact with these tools? When do users want or not want to use the algorithmic tools made available to them? How does reliance on these tools affect how communities evolve? Following [3], we know that volunteer moderators develop complex, community-specific rules as their communities evolve, but the literature lacks a generalizable, cross-platform model for how these decisions are made over the life cycle of a community.

The overarching theme emerging from existing community moderation literature is that communities evolve over time as a result of rule-breaking, rule-making, and rule-enforcement [28, p. 158-169]. Rule-breakers often include malicious outsiders, spammers, or trolls but sometimes also regular users who misunderstand rules or get carried away. It is important to note that, while these approaches are content-focused, users are not passive participants in the moderation cycle; they actively monitor and react to both algorithmic and human moderation decisions to gauge what is appropriate or not. Some even do this to evade detection for content that they know will violate the rules [5, 6].

Grimmelman defines moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." [9, p. 52]. This definition matches well with our findings in this study; our interviewees mod-

erated communities based around computer games, board games, art, memes, cars, sports, pets, and learning. They acted as arbiters, governors, community managers, teachers, role models, curators, and enforcers. Modern online communities look much like their earlier incarnations both in the misbehaviors they handle and the strategies they use to guide the community, but they exist in a new context: working alongside - and sometimes at odds with - platform-driven algorithmic moderation. The present work aims to provide an in-depth examination of user-driven moderation to shed new light on the development and decision making processes exhibited by volunteer moderators across diverse platforms.

3.4 Methods

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

In the second phase of this project we interviewed 21 Reddit moderators[5]. In

[5]One Reddit moderator interview, R14, could not be transcribed due to audio issues, so is not

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

Third, we interviewed 15 Facebook Group moderators[6] during Fall 2017 through Spring 2018. We focused on the same types of groups as on Reddit, using keywords related to the subreddits from which we had previously interviewed moderators (e.g., "cars" and "memes"), with the goal of finding comparable communities on all three platforms. We selected Facebook groups of various sizes (500-70,000 members) and messaged recently active moderators for interviews.

Once all interviews were completed and transcribed, we "winnowed" the text into chunks for coding following the procedure established in Creswell [2, p. 86-89, 184-185]. Each chunk contained a single idea and varied in length from a few words to three sentences. Our final dataset contained 1,877 chunks of text. One rater assigned codes to a subset of these chunks. First, low-level themes were identified, and then these themes were abstracted to higher levels of generality to produce a comprehensive codebook.

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

included in counts in Appendix C

[6]Our sample included both what Facebook calls "moderators" and "admins", both of which are types of moderators as defined by [9, p. 42] with the latter having more permissions. For the sake of simplicity we refer to both as "moderators" here and use the term "platform administrators" to refer exclusively to employees of the respective companies.

Disagreements were resolved through discussion.

This sample is not a random sample of moderators on these sites. Female and LGBTQ+ moderators are intentionally over-represented in our sample, especially on Twitch, in order to gather a more diverse set of experiences. Experiences of under-represented moderators are especially important to capture on Twitch because the video-based structure of the platform makes their identity more salient to community members, and because of the prevalence of sexism in game-related media [4].

3.5 The Moderator Engagement Model of Community Development

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

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

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

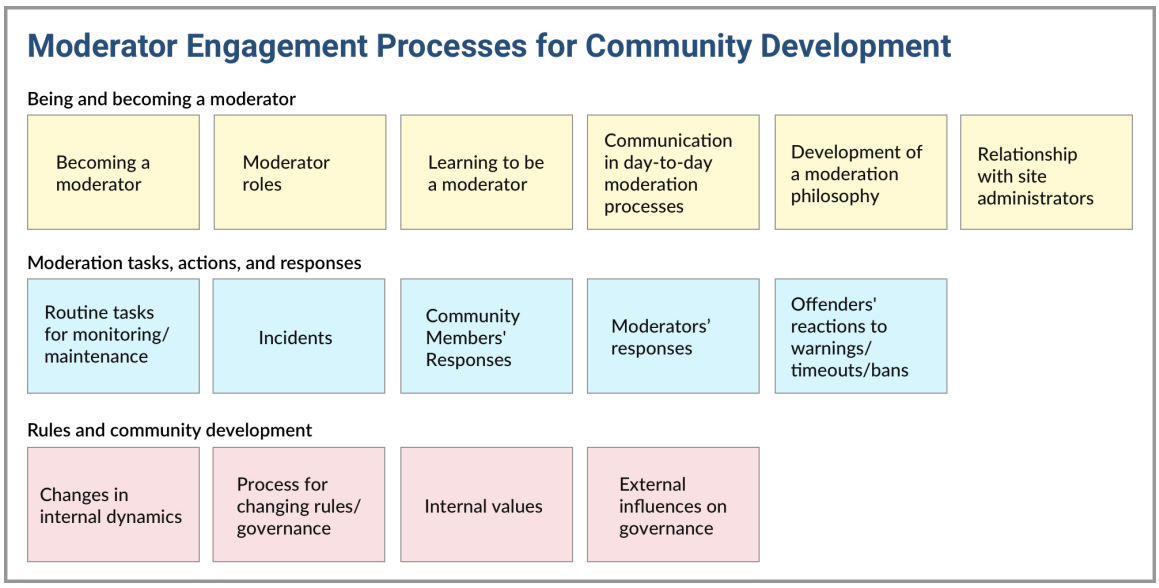

Figure 3-1: Moderator Engagement Model of Community Development

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

- R21 has been moderating a sports subreddit for three years. Instead of being given formal guidelines, R21 was told by more senior moderators to moderate however R21 feels is best for the subreddit. The moderation team is particularly wary of violating a Reddit rule against posting personal information, so R21 spends time warning users about this. R21 has worked with Reddit administrators before, so when a user repeatedly violated the subreddit's rules by making new accounts after being banned, R21 reached out to them for help. In this example, specific incidents led to community-specific rule enforcement styles and eventual engagement with platform administrators.

- T18 streams on Twitch to an average of several hundred concurrent viewers. T18 has no hard-line rules about moderation. Instead, T18 focuses on engag-

ing viewers who share different opinions and only removes spam bots. Though Twitch has added new moderation tools in recent years, T18 feels that these tools have no place in his community because moderation tools stifle conversation and inhibit safe and productive community growth. In this third example, a particular moderation philosophy adheres to social rather than technical forms of moderation, which are intended to help the community grow in a particular way.

3.5.1 Being and becoming a moderator

The ways in which individuals become moderators are similar across platforms and are largely driven by discussions within moderating teams. Platform administrators are occasionally involved, but these instances are rare and often opaque to the moderators. The steps shown in column one of Figure 3-2 match steps from the overall process diagram, Figure 3-1. The second column shows themes and variants within each step. See Appendix C for counts of each code.

Becoming a moderator

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

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

Users could also become moderators because they had an existing relationship with current moderators of a community. This practice is more common on Facebook

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

Figure 3-2: Steps and variants in Being and Becoming a Moderator process [$\kappa = 0.89$]

than either of the other two sites, likely because Facebook is a social network based on real-life relationships and allows for "friending." Female moderators in smaller Twitch communities noted the value of having friends moderate for them, particularly ones who understood the sexism they faced. This matches recent work on "friendsourced moderation" [18].

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

Role differentiation

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

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

Though differentiated authority was reported across all three platforms, the way it manifested depended on the platform's design. On Twitch, role differentiation is explicit in every community: the channel owner (usually the streamer) has ultimate authority in their own community. Facebook offers different levels of power through the "administrator" and "moderator" designations. Reddit has less obvious titles, but more senior moderators can add new moderators with limited permissions. Each of these design decisions facilitates a slightly different type of status differentiation.

Learning to be a moderator

Formal education of new moderators was strikingly rare on both Twitch and Reddit. More than two thirds of moderators learned based on a combination of understanding

the community's values and simply learning by doing.

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

Fewer than a third of our interviewees had any formal onboarding conversation about what was expected of them, and even fewer were given formal guidelines. Of the three platforms, these conversations were most common on Facebook; moderation processes are typically invisible to casual Facebook users who are not involved in Groups [21], so Facebook users may have simply needed more of an introduction to them:

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

Communication between moderators and Development of a moderation philosophy

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

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

This was less common on Twitch. Due to the synchronous nature of conversations on Twitch, moderation decisions need to happen immediately. However, Twitch moderators communicated in other ways, including "mod meetings" where incidents and rules were discussed outside of streaming hours. Some moderators indicated that there was

rarely any communication at all, and though they did not usually see this as a problem, they were typically from groups that were either homogeneous or stagnant. A lack of conversation often indicated a lack of growth. One Twitch moderator (T18) noted that this was the reason he avoided algorithmically-driven moderation; ongoing discussions about acceptable behaviors helped both him and his community grow, and automating these decisions could remove opportunities to have these conversations.

Over time, moderators developed a philosophy through conversations with other moderators and engagement with users. Four different philosophies emerged in these interviews. A number of moderators stated that they felt it was important for them to be very present in the community by setting an example and engaging directly. Other moderators saw themselves as "police", tasked with identifying and punishing offenders in order to keep the community civil. A third group of moderators sought to maintain a neutral perspective. They felt suspected bias could affect their credibility or hinder open discussion. They stepped in when conversations got out of hand, but did not take stances themselves. And finally, a small group of moderators believed in a hands-off approach, interfering as little as possible even when discussions devolved into personal insults:

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

This fourth philosophy was associated with a strongly anti-"censorship" ethos [19], with the idea that debate and discussion naturally surfaces better ideas and greater understanding.

Relationship with site administrators

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

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

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

Prior work [20] has explored moderators' relationships with platform administrators, noting moderators' need to maintain face. We found few examples of this; most of our interviewees felt no connection to administrators and believed their communities would never host behaviors that would attract administrator attention.

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

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

Beyond these relationships, several moderators reported "mod burnout" in which moderators became exhausted by the amount of work and exposure to offensive content. Several moderators recalled traumatic experiences like threats and harassment, some cases of which even followed them offline. To many moderators, moderation

is equivalent to a second job where they work for the benefit of the platform and are rewarded only with the satisfaction of helping shape a community that they care about [20].

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

3.5.2 Moderation tasks, actions, and responses

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

Routine tasks for monitoring/maintenance

The asynchronous nature of Facebook Groups and Reddit encourages moderators to develop proactive strategies for preventing misbehavior. This reduces the amount of work they have when responding to incidents after they happen. Almost all Facebook Groups interviewees reported that they spend a significant amount of time reviewing join requests, a feature that does not exist on either Twitch or public subreddits. Join requests give moderators the ability to accept or deny prospective members,

2. MODERATION TASKS, ACTIONS, AND RESPONSES

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

Figure 3-3: Steps and variants in Moderation Tasks, Actions, and Responses process [$\kappa = 0.70$]

which moderators reported was effective at reducing spam but created an additional workload for them. While most groups only filtered to prevent spam accounts, some groups used filtering to curate members that had similar values, interests, or political orientations. Facebook Groups moderators also spent significant amounts of time monitoring potentially controversial threads and stepping in to preclude conflicts. Prior work has shown that there are detectable patterns of behavior that precede misbehavior [16], and Facebook moderators are likely attuned to these.

A number of moderators, particularly on Twitch and Facebook, felt it was their duty to engage regularly with their community. Some Facebook moderators contributed content on a regular basis to keep discussion flowing, and Twitch moderators regularly welcomed new members and actively answered questions. Some framed these behaviors as setting a positive example. Prior work has found that users imitate authority figures' behavior online at significant rates, and moderators took advantage of this [25]. Reddit moderators were generally more hands-off in this sense.

Incidents

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

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

ation challenges, and often assumed that users who misbehaved were outsiders or newcomers.

Community members' responses

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

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

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

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

Moderators' responses

When proactive approaches and general community rebukes failed, moderators stepped in to address misbehavior directly. In general, moderators' responses to offenses began as light, verbal warnings and escalated into increasingly technical restrictions after repeated occurrences such as stronger warnings with a temporary mute and, eventually, a permanent ban. Exceptions included spam, egregious content, or suspected non-human accounts - these offenses typically warranted immediate bans. Virtually all moderators on Twitch and Facebook reported warnings as the first step in many

cases, especially if offenses were mild and unintentional. As in Slovak [27], moderators saw themselves as arbiters of the rules but also as teachers helping users learn how to behave.

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

Nearly all moderators mentioned using timeouts, bans, or equivalents, though eagerness to use them varied. Communities with more laissez-faire ideologies used these only for egregious offenses, while communities intended to be safe spaces were usually quicker to use them. While Twitch moderators relied extensively on warnings, they rarely issued explanations of punishments after the fact unless privately contacted by a user. This is likely because of the synchronicity of conversation on Twitch. On the other hand, Facebook and Reddit moderators frequently explained why they punished users or removed content, especially for posts that were left "hidden" until changes to the content were made to fit moderator approval.

There were three main platform-provided, algorithmically-based moderation tools used by moderators in this study. Moderators on Reddit relied substantially on the built-in AutoModerator to parse, flag, or remove suspicious posts. Despite its utility, AutoModerator sometimes created additional work for moderators because they had to manually approve posts mistakenly "caught" by the bot. Twitch moderators relied less on site-provided automated tools with the exception of the emerging "AutoMod" tool, [7] and Facebook Groups moderators relied the least on automated tools to parse posts for them, though some made use of built-in filters in the user approval process where potential spam accounts were marked for review.

Many Twitch and Reddit moderators looked beyond site-provided tools and used free user-developed bots; some even created their own. Most of these indepdently managed tools were simple filters that flagged custom words; moderators preferred

[7] Note that our interviews took place largely prior to the widespread proliferation of Twitch's "AutoMod"

to engage verbally with human offenders in more nuanced situations. Beyond bots, the most commonly used tools were chat logs, post histories, and ban logs. Facebook Groups moderators were least likely to integrate custom tools because of third-party developer restrictions on the site.

Moderators we interviewed were happy to have tools that deal with the most obviously unwanted content, such as links to malware or pornography, but they have a strong preference to make the hard decisions themselves. This stands in contrast to prior work on preferences for algorithms in other contexts. When exploring users' preferences for algorithmically-driven tools in medical decision-making processes, Yang [30] found that doctors felt no need for these tools for most of the easy decisions they made because they felt confident in their decision-making ability, preferring support instead on the harder decisions. The difference between these cases likely results from the importance of continuously-evolving community values in decisions made by moderators. They noted both the importance of their ability to make context-specific judgments and also the impact on the community's development that their decisions have as justification for reserving these decisions for human judgment.

Offenders' reactions to warnings/timeouts/bans

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

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

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

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

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

Top-down approaches in commercial content moderation are designed to minimize subjectivity in moderation work [8, p. 111-114]. According to our interviewees, however, it is this very subjectivity that helps communities develop over time. The values that moderators brought to their commmunities and the ways that these values changed as a result of interactions with users were core to community growth. In cases where decisions were made by algorithms rather than moderators, these moderators might not have the same opportunities to grow.

3.5.3 Rules and community development

Communities and their rules develop over time through reactions to short-term events or transitions, as shown in Figure 3-4. As rules and norms develop, they drive subsequent moderation decisions that shape community identity. These decisions are almost exclusively made by moderators, either by an executive "head" moderator or a group consensus. Occasionally, moderators solicited feedback from their communities, but this was rare. General community members are rarely given a say in the final outcome.

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

Figure 3-4: Steps and variants in Rules and Community Development process [$\kappa = 0.85$]

Changes in internal dynamics

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

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

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

Rules might also change in response to unusual internal or external events. One

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

Process for changing rules

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

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

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

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

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

> "If we see a lot of people complaining about a rule we re-evaluate it. Back before when [the community] was a lot smaller it used to be more lenient about generic posts, like âĂIJheres a photo of me at the gameâĂİ or âĂIJheres a photo of me with a playerâĂİ but we kind of put a stop to those just because [the community felt] it wasn't as interesting to discussion. (R19)"

External and internal influences

We identified two other sources of influence in making these decisions: external influences (such as sitewide policies) and internal influences (such as personal values). Influence of site rules and content policies on community rules was rarely present, likely due to the vague and distant nature of these policies [7, 23]. Reddit's official policies did serve as a minimum standard for behavior in subreddits, as failure to comply could lead to shutdown of the community. While Reddit moderators had little to

no contact with site admins, they did note that specific policies impacted how they moderated, with the most common being Reddit's policy against revealing personal information. Moderators on Twitch and Facebook usually did not pay attention to content policies and, in many cases, did not know what they were.

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

3.6 Conclusion

This article outlines three processes through which moderator engagement guides the development of online communities: becoming and developing as a moderator, handling misbehavior, and developing rules for the community. It contributes to the growing discussion surrounding management of behavior in online spaces by documenting the social nuances of moderation that disappear when moderation is delegated to commercial content moderators or automated algorithms. These findings emphasize the need for a closer look at social, user-driven models of moderation. The communities described by just the 56 moderators in this study are comprised of more than five million total members, many of whom find valuable connections, relationships, and meaning in these spaces. While recent work has suggested that social media companies may be the "New Governors" of the digital age [15], it is important to remember that this centralized aggregation of power is not necessary for meaningful online socialization to occur. Users can be very effective at self-governing when given the tools to do so, and this experience in itself can be meaningful.

Both community-based moderation and commercial content moderation have clear

drawbacks. As Gillespie [8] notes, there are a variety of challenges in commercial content moderation that platforms must address if they are to be the "custodians" of the public sphere. They must balance intervention with protection of users' rights to speak; they must make moderation decisions constantly; and they must maintain the appearance of fairness and objectivity. Though volunteer user moderators also sometimes struggle with transparency and fairness, they are much better equipped to understand the context surrounding issues in their communities. Moderators engage personally in dealing with a variety of nuanced problems, guide conversation in positive directions, and are a regular, stable presence in their communities. Moderation algorithms and commercial content moderators are unlikely to be able to contribute to these community attributes in the same way.

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

Prior research has explored ways to make commercial content moderation more effective because of its presumed scalability, but it is important to remember that online communities have been self-governing imperfectly but effectively since the beginning of the social web. Many of the challenges in scaling moderation, such as inability to consider context of each potential violation, come from decisions to structure social platforms like Facebook and Twitter as networks rather than a series of self-governing communities. Even within Facebook itself, we find that characteristics of user-driven moderation within Groups are largely consistent with the characteristics within Twitch and Reddit communities; Facebook Groups offer an opportunity to test what a more community-based Facebook might look like and whether this model might be able to avoid some of the pitfalls of Facebook's default commercial content moderation approach.

There are many ways to balance algorithmic and user-driven models of governance, and each implementation has different implications for communities online.

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

We offer four additional considerations for future design in light of these findings. First, platforms should work to develop and improve tools that allow moderators to focus their attention where it is needed. For example, Facebook and Reddit moderators currently rely extensively on flags, but predictive suggestions for threads or discussions that might soon devolve could be useful as a complement to flags [16]. Second, platforms should consider features that encourage positive behaviors in meaningful ways. While tools for dealing with misbehavior are common, tools for encouraging meaningfulness are limited at best (e.g., Reddit gold, Facebook reactions). Third, platforms might consider developing features that allow all users to get involved in self-governance if they so choose. Moderators in our study made decisions by executive fiat usually without community input, and could not be removed from their positions except in some cases by other moderators. This model of governance has been common across online communities since the early social web, but other models of community governance may be possible. Finally, platforms should consider where scaffolded user-driven moderation might serve communities better than algorithmic or company-driven moderation, and how social features might be designed to facilitate user-driven moderation. For example, when is a network a better design choice

than a set of communities, and vice versa? What would Twitter and Facebook look like if they were structured primarily around communities rather than networks?

Meaningfulness in online spaces emerges from nuanced social interactions, both positive and negative, and volunteer moderators are at the core of these interactions. Through this analysis, we document how moderators help such communities grow, evolve, and become more meaningful for their members as they work through the challenges that come from engaging with new media. We find that, while moderators did in certain cases make use of algorithmic moderation tools, they always sought to give the important parts of their jobs careful human attention and contextually-informed judgment. Future tools should support moderators in finding time and energy to focus on the tasks that they find meaningful and that help their communities grow.[8]

Bibliography

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

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

[3] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. Reddit Rules! Characterizing an Ecosystem of Governance. In *International AAAI Conference on Web and Social Media*, 2018.

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

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

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

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

[7] Tarleton Gillespie. The Politics of 'Platforms'. *New Media & Society*, pages 1–3, 2010.

[8] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* Yale University Press, 2018.

[9] James Grimmelmann. The Virtues of Moderation. *Yale Journal of Law and Technology*, 17(1), 2015.

[10] William A. Hamilton, Oliver Garretson, and Andruid Kerne. Streaming on twitch: Fostering participatory communities of play within live mixed media. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1315–1324, New York, NY, USA, 2014. ACM.

[11] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society*, 18(5):371–384, 2002.

[12] Zorah Hilvert-Bruce, James T Neill, Max Sjöblom, and Juho Hamari. Social motivations of live-streaming viewer engagement on Twitch. *Computers in Human Behavior*, 84:58–67, 2018.

[13] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):12, 2018.

[14] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design*, 2012.

[15] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 2018.

[16] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *International AAAI Conference on Web and Social Media*, volume 91, pages 181–190, 2018.

[17] Claudia Lo. When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators. Master's thesis, Massachusetts Institute of Technology, 2018.

[18] Kaitlin Mahar, Amy X Zhang, and David Karger. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *CHI '18*, Montreal, QC, Canada, 2018. ACM.

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

[20] J Nathan Matias. The Civic Labor of Online Moderators. In *Internet Politics and Policy Conference, Oxford, United Kingdom*, 2016.

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

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

[23] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Group '16*, pages 369–374, Sanibel Island, FL, USA, 2016. ACM.

[24] Sarah T Roberts. Commercial Content Moderation: Digital Laborers' Dirty Work. In S U Noble and Tynes B., editors, *The Intersectional Internet: Race, Sex, Class and Culture Online*, pages 147–160. Peter Lang Digital Formations series, 2016.

[25] Joseph Seering, Robert E Kraut, and Laura Dabbish. Shaping Pro and Anti - Social Behavior on Twitch Through Moderation and Example - Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017.

[26] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society*, 2019.

[27] Petr Slovak, Katie Salen, Stephanie Ta, and Geraldine Fitzpatrick. Mediating Conflicts in Minecraft: Empowering Learning in Online Multiplayer Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 595:1–13. ACM, 2018.

[28] Janet Sternberg. *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet.* Rowman & Littlefield, 2012.

[29] Donghee Yvette Wohn, Guo Freeman, and Caitlin McLaughlin. Explaining viewers' emotional, instrumental, and financial support provision for live streamers. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 474:1–13. ACM, 2018.

[30] Qian Yang. The role of design in creating machine-learning-enhanced user experience. In *2017 AAAI Spring Symposium Series*, pages 406–411, 2017.

Chapter 4

Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting

4.1 Introduction

In thriving online communities, a rough consensus generally emerges about norms, i.e. the range of acceptable behaviors.[1] Norms of appropriate behavior vary substantially across communities. Personal insults may be the primary way to interact in one community, but may be frowned upon in another. Wikipedia expects writers to adopt a neutral point of view when writing articles, while the Huffington Post expects guest bloggers to express a viewpoint. PsychCentral.com, a site with more than 150 health support communities, prohibits conducting any type of research on the site for publication or educational purposes [2]. Snapchat users share mundane everyday moments in solidarity with friends [9].

Communities can use formal and informal methods to enforce standards of appropriate behavior [21], including explicit rules, reputation systems that provide incen-

[1] This chapter draws primarily from my paper, "Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting" [30].

[2] PsychCentral. 2008. Terms of Use, from `http://psychcentral.com/about/terms.htm`. Retrieved 8 Mar 2008.

tives for people to act appropriately, methods to report inappropriate behavior, and algorithms that automatically remove offending behavior. While these approaches help deal with misbehavior, anti-normative behavior is still a substantial problem on a variety of platforms [5]. This paper explores how moderation tools and imitation effects can address norm violations. We explore the effectiveness of both proactive (preventive) and reactive (punitive) moderation tools, where proactive tools prevent certain behaviors while reactive tools punish users after-the-fact for engaging in them. We also explore the potential for authority figures to shift culture by modeling positive behaviors other users can emulate.

This paper builds on previous research on imitation and deterrence and applies these concepts to moderation in a pseudonymous, ephemeral environment. First, building on lessons from theories of imitation [6, 10, 11, 12, 25, 35], we show that text chat behaviors on Twitch (twitch.tv), a video streaming platform, are contagious, and that anti-social and pro-social behaviors spread differently. We show that the behavior of individuals with authority in and commitment to a particular channel has greater influence over the behavior of others, and that "outsider" users without specific status in a channel have no impact. Second, drawing from Deterrence Theory [13, 17, 24, 33], we demonstrate that different approaches to moderation can reduce the spread of different types of behavior. Setting a chatroom to a restrictive mode reduces the frequency of spam overall while having no substantial effect on other types of behavior. Banning a user after they post a message of a certain type lowers rates of that type of message in subsequent posts.

4.2 Background

This paper explores two approaches to influencing behaviors on Twitch: behavioral imitation, where observing one type of behaviors encourages observers to behave in the same way, and deterrence, where threat of punishment or enacted punishment causes users to change their behavior. This section explores relevant findings in the literature in each of these mechanisms, and considers what each theory would predict

for the Twitch context.

4.2.1 Imitation and Conformity

Several related processes lead to the spread of behavior from one person to another. Theories of imitation discuss how individuals learn to behave and think in ways that are similar to their peers. Theories of obedience and conformity describe the power of authority figures to influence behavior.

Imitation occurs in two primary phases. The initial effect of observing others' behaviors, as described by Wheeler, is in helping the individuals conceive of those behaviors as possible courses of action [35] when they might not otherwise have considered them. Here we will refer to this as the conception effect. Second, when choosing from a set of possible behaviors, individuals are more likely to choose the behavior that their peers prefer or which they perceive to be in accordance with social norms [6, 10, 12, 25]. The more peers observed acting in a certain way, the more likely the individual is to do so as well, independent of whether such an action is a good idea.

Studies have found the presence of imitation effects in a variety of different contexts, from the mimicry of non-verbal behavior and language [10], to expressions of attitude and beliefs [7, 32] to speeding [29], to copycat suicide [27].

Social learning theory contributes an additional related perspective on mechanisms involved in the adoption of deviant and conforming behavior [4, 16]. This framework explains a variety of what Akers describes as deviant behaviors including adolescent marijuana usage [2], teenage cigarette smoking [1], and alcohol use among the elderly [3].

We hypothesize that behavior on Twitch can be explained by these same types of imitation effects. Rare behaviors will be imitated at particularly high rates as a result of the conception effect [35]; while on Twitch certain behaviors such as spam are very common and users need no reminder that they are possible courses of action, reminders to be polite and kind will have more influence due to the relative rarity of those types of behaviors.

H1: An instance of a given behavior on Twitch will increase the likelihood of other users engaging in that behavior, and behaviors that are more rare will be imitated at greater rates.

Literature on conformity and obedience to authority supplements the perspective provided by imitation theories. Milgram's classic shock experiments showed how powerful the effect of an authority figure can be on overriding individual tendencies [23]. Nurses were willing to follow alarmingly bad directions if encouraged to do so by an authoritative doctor [18]. Psychological journal articles were much more likely to be accepted for publication if they had well-known researchers' names attached [26]. Overall, people are more likely to follow the example or take the word of others who have explicit authority.

Imitation has also been documented in Human-Computer Interaction literature [8, 20, 28, 36]. Bakshy et al. [8] found that Facebook users were significantly more likely to share a link if their friends had shared it. This effect was limited to a brief period of time after they saw their friends' posts. Zhu, Kraut, and Kittur also found significant imitation and conformity effects among Wikipedia workers collaborating on projects, though this effect varied significantly based on level of users' identification with the group [36].

In each of the cases discussed above, participants were more likely to do something that they would otherwise be uncomfortable doing, or that under other circumstances they would find reprehensible. From conformity theories we draw the hypothesis that users with authority or status will be more likely to be emulated:

H2: Users with more authority or status within a community will be imitated with greater frequency.

There are four different status badges that users on Twitch can have in a given channel, each of which is denoted by a specific badge that appears next to their name when they post in chat. Channel owners are users who broadcast the content on the channel. They have the highest status, and have ultimate authority over the various moderation tools. Moderators are users designated by the channel owner to enforce behavioral standards through bans and chat moderation modes. They have

authority and status within the community. Subscribers are users who have paid a monthly fee of approximately $5, of which part goes to the channel owner and part to Twitch, to gain special privileges in a channel and to support the broadcaster. Subscription is only available as an option in sufficiently large or well-known channels, but through subscription to a channel a user may quickly progress to a higher status by way of demonstrated commitment to the channel [15]. Twitch Turbo users pay approximately $9 per month to gain small benefits across the whole site, including removal of ads. A Twitch Turbo badge is a mark of some status, but not status within the ingroup. In this study, we will refer to users with none of these statuses as regular users. Note that we do not include channel owners' chat messages in our analyses because channel owners usually communicate with viewers by speaking directly as part of their broadcast.

Here we use these user status categories as representative of different levels of authority and commitment to a channel. Per Hogg's social attraction hypothesis [19], ingroup members are more liked and thus more influential because they are perceived as conforming to a positive ingroup prototype. Here, users who like or want to be part of a particular channel community see the behavior of subscribers as clear examples of ingroup norms, as these subscribers have demonstrated explicit loyalty to the channel. Moderators, through both their additional abilities and their status as a favorite of the channel owner, exemplify Hogg's prototypical leaders: they have disproportionate power to determine standards of conduct, define identity, and organize and guide discussion. In contrast, Twitch Turbo users have a literal status icon next to their names, but this status is not specific to the channel. Within the channel they are as much outsiders as regular users if not more so.

4.2.2 Deterrence Theory

Where imitation and social learning theories focus primarily on the ways behaviors are learned from peer groups and networks, Deterrence Theory focuses specifically on the impact of punishment on deterring certain types of behaviors, and is used here as a reference for understanding the impact of certain moderation strategies on behaviors

[13]. Deterrence Theory distinguishes between general and specific deterrence, where specific deterrence is defined as the impact of punitive actions on individuals upon which they are enforced, and general deterrence is the impact of the threat of such action on uninvolved observers. Deterrence theory has also been conceptualized as indirect vs direct experiences with punishment [33].

The theory of general deterrence suggests that the threat of arrest and punishment may deter criminals from committing crimes, and that different levels of certainty and severity of punishment will affect the effectiveness of this deterrence. This effect has been demonstrated in online contexts [17]. Nagin [24] identifies several methods for studying deterrence in the wild, including interrupted time-series studies, ecological studies, and perceptual studies. This current study uses the interrupted time-series method, a quasi-experimental method that looks at differences in behaviors immediately before and after an intervention. In this case we look at chat behaviors prior to and subsequent to a deterrence event that we hypothesize will affect these behaviors and compare the effects of deterrence on different categories of behaviors. By looking at thousands of instances of the intervention, our procedure guards against many confounds associated with interrupted time series methods, e.g. other historical events occurring simultaneously with the intervention.

The Twitch platform offers several tools to help moderate offensive behaviors. Broadcasters and moderators can ban users directly for variable lengths of time in response to an offensive or inappropriate message. Various third party chat-moderation bots can also be installed to automatically ban users who post certain specified types of content. Channel chat moderation modes can be enabled by channel owners or moderators proactively to prevent certain types of posting behavior.

While Nagin discusses the challenges of understanding the connections between perceptions of abstract policies and impact on an individual's behavior, the immediacy of punishment on Twitch helps avoid this particular pitfall. Where real-world legislators may seem distant from the corporeal behavior that they regulate, the "policy-makers" on Twitch are often literally visible to their audience. Furthermore, punishments in the form of bans are relatively common and are visible to all users,

so participants have clear and direct evidence of what is and is not considered appropriate behavior. In many cases they are often even told directly by chat moderation bots or human moderators what specific behavior caused the ban.

We hypothesize the presence of a generalized preventative effect from proactive moderation techniques on Twitch, which take the form of chat moderation modes. Channel chat moderation modes are tools available on all channels that restrict users' posting behaviors. The three modes we explored are subscribers-only mode, where only subscribers to the channel may chat, slow-mode, where users have to wait a specified amount of time between sending messages, and R9k-beta mode, where users are prohibited from posting lengthy content that has already been posted. Whereas traditional bans are imposed in response to messages, these modes prevent messages from being posted at all except under the designated circumstances. These modes cannot be customized to target different types of unwanted behavior, beyond the choice to enable or disable them in response to different chatroom conditions. As such, channel owners and moderators cannot directly customize their usage to encourage particular behaviors; these modes only serve to make it more difficult to engage in specific anti-social behaviors, namely spam. Because of this, they will have a generalized preventative effect on anti-social behavior, but will not exhibit classic generalized deterrent effects and will not directly encourage pro-social behavior.

H3: When chat moderation modes are enabled, the frequency of spam will decrease, but the frequency of other more prosocial behaviors will not increase.

In contrast to the limited customizability of chat moderation modes, bans are completely customizable; the owner of a particular channel chatroom and the moderators they designate choose which users to ban and for how long, and what settings to use on moderation bots that ban users by proxy. Bans, like chat moderation modes, will have a deterrent effect, but this effect will be more flexible and will deter different types of behavior depending on how bans are applied.

H4: When a particular type of behavior is banned, the subsequent messages will have a lower frequency of this type of behavior.

While we have framed this hypothesis in terms of deterrence theory, imitation the-

ory makes a similar prediction, although for a different reason. Banning a particular type of content removes it from the view of other users, who then simply may not conceive of it as possible behavior [35].

4.3 Procedure

This study involved the collection and analysis of approximately 21 million messages sent to channel chatrooms on Twitch over the course of one week in early March 2016. In this section, we first describe the method for gathering data and then describe the four analyses performed on this data. The first analysis demonstrates clear imitation effects for three different types of behavior that we studied: spam, questions, and smiles. The second analysis shows that the status and authority of users within the ingroup affects the amount that they are imitated, and that the effects vary across different behaviors. The third analysis shows differential results from reactive and proactive approaches to moderation. Finally, the fourth analysis takes a different approach and attempts to determine the duration of the impact of these effects by looking at the strength of the imitation effect over time.

Twitch, described in more detail in the previous chapter, is an ideal platform on which to study the effects of different moderation techniques. With thousands of channels with different numbers of viewers, different approaches to moderation, and different behavioral norms, analysis of chatroom data allows for a better understanding of what works to stop the spread of undesired behaviors. Behavioral imitation is very visible on Twitch. For example, users in larger channels often engage in spamming of "copypasta," which are long, often-nonsensical messages with many emotes that users copy and paste into a chat repeatedly. While such behavior may be desirable on some channels and in some cases it can even be compared to the type of cheering that happens at sporting events[3], many broadcasters prefer to keep their chat rooms civil and thus seek tools that will stop the spread of such behavior when

[3] https://web.archive.org/web/20190723181512/https://speakerdeck.com/drewww/past-and-future-of-game-spectating

it appears.

4.3.1 Data collection

In this study we identified three categories of behavior that represented different modes of interaction on Twitch in order to understand how they spread and persisted or disappeared in response to moderation. First, we defined "spam" using the default settings on one of the most widely used Twitch chat moderation bots as an example of anti-normative behavior. By this definition, spam messages were those with a large number of emotes, capital letters, or symbols. Second, we identified conversational messages, where users ask questions of the broadcaster or each other as an example of neutral behavior. These are messages that end with a question mark. Third, we identified messages with positive emotions, which in this case are defined as those with positive smile emoticons, as an example of pro-social behavior. Note that we classify large numbers of emotes unique to Twitch as spam in accordance with common moderation bot rules and with the understanding that many of these emotes are used for trolling [14], but we found in our analysis that single uses of traditional smiley-face emoticons were almost always positive or only lightly teasing.

Note that for this study we focused on Twitch-wide emotes instead of channel-specific emotes; the Twitch-wide emotes are much more widely-used, and meaning can be generalized across channels. "Kappa", the most used emote on Twitch, is posted upwards of one million times per day across Twitch [4].

We selected these categories as examples of a range of behaviors, but it is important to note that in many channels on Twitch "spammy" messages are tolerated or even encouraged. Table 1 shows the three types of messages that were analyzed and their overall frequency in the dataset.

Contagion of the sort that we discuss here can take a variety of forms. For example, spam contagion might be started by a small number of users to disrupt a particular conversation:

[4] `https://web.archive.org/web/20190624064939/https://fivethirtyeight.com/features/why-a-former-twitch-employee-has-one-of-the-most-reproduced-faces-ever/`

Type	Criteria for inclusion	Action Valence	Overall Frequency as % of Messages
Spam	Many emotes, capital letters, or symbols	Anti-normative	14.8%
Questions	Ends with "?"	Neutral	4.7%
Smiles	Contains ":)", ":D", ":P", or ";)"	Pro-social	0.9%

Table 1: Categories of Messages

USER1: go [play] constructed please!

USER2: constructed is usually boring to me

USER3: I wouldn't mind watching you do maybe a couple of hours of constructed a week, just for a change.

USER4: are you ever going to play SMITE again @[streamer]?

USER5: VapeNation

USER6: V/\Vape Nayshon!

USER7: NapeVation /\V

USER8: why are all u idiots spamming vapenation

USER9: I love waiting for a comment to get [banned] haha

USER10: VapeNation

USER11: VapeNation V/\

USER12: VapeNation V/\

In this example, users were having a conversation with the streamer about what the streamer should do next when they were interrupted by other users spamming variants of "VapeNation", a popular meme-phrase referring to smoking with a vaporizer. Regular users expressed irritation about this disruption and requested bans for the spammers, but in this case the moderators did not intervene and the spamming continued for another five minutes before it tapered off.

Imitation of question-asking behaviors might be encouraged by a perceived likeli-

hood to get a response from the streamer or a moderator:

> USER1: @[streamer] what do u think about jax jungle ?
> USER2: @[streamer] Any chance we can see Skarner jng of the team comps permit it?
> USER3: @[streamer] What do you think of Aurelion Sol? What's the best role to play him/her in?
> USER4: @[streamer] just got out of a game with morgana..bot.we were ok...adc was feded..but mid and jungle fed...so what do i do in such occasions??

The streamer in this example, who is a highly ranked player of the popular game League of Legends, probably responded verbally in the stream to the first question asked here, so other users decided that it was a good time for them to ask questions as well. Note that this mechanism is imitation but is mediated by the actions of a third party, which in this case is the streamer.

Finally, imitation of smiles often comes from positive exchanges between users in the chat, and can spread to other interactions as well:

> USER1: hiii @[USER2] :D
> USER2: @[USER1], hey :)
> USER1: :D
> USER2: hi chat :)

Here one user greets another in a friendly way, and this user responds positively and in turn greets the other users.

While most Twitch users only see the graphical interface for the channel chatrooms (see Figure 1), the underlying structure of the chats is based on IRC (Internet-Relay-Chat) protocols. We created an IRC "chatbot" in Python that uses the Socket module for the purpose of collecting messages and relevant metadata from channels across

Twitch. The data collection process involved three steps: determining which channels to scrape, creating the data collection scripts, and running the scripts over the period of one week to create the dataset. This chat scraping was done with the permission of data science staff at Twitch.

This study used sample a of roughly 600 Twitch English-language channels stratified by size. In order to select channels for analysis, we created ten strata of channels, based on their number of viewers, where each stratum contained approximately the same number of viewers. The first group, which contained the largest channels, had the fewest channels overall, while each successive group had more channels with fewer viewers. We randomly selected 100 channels from each of these groups with the exception of the first group and the third group, which contained only 81 and 46 channels total.

We ran the message collection script for nine days in early March 2016. During this time period, every message sent to one of the channels in our sample was recorded. Overall, approximately 21 million messages were collected from English-speaking channels.

4.3.2 Features of the dataset

Each message in the dataset was tagged with several features. These included the channel name, username, timestamp, and message as well as whether the user who sent the message was a moderator, subscriber, or Twitch Turbo user. We removed usernames from our dataset prior to analysis. We also tracked whether each channel had a channel chat moderation mode set while any given message was sent, which would have influenced the types of messages that could be sent. While messages were not tagged directly tagged as being banned, bans are user-specific not message-specific, we inferred that a specific message was banned if it was the most recent message sent by a user who was banned, under the assumption that users are most frequently banned in response to the most recent message they sent prior to their ban.

After the data was collected, we tagged each message with a variety of charac-

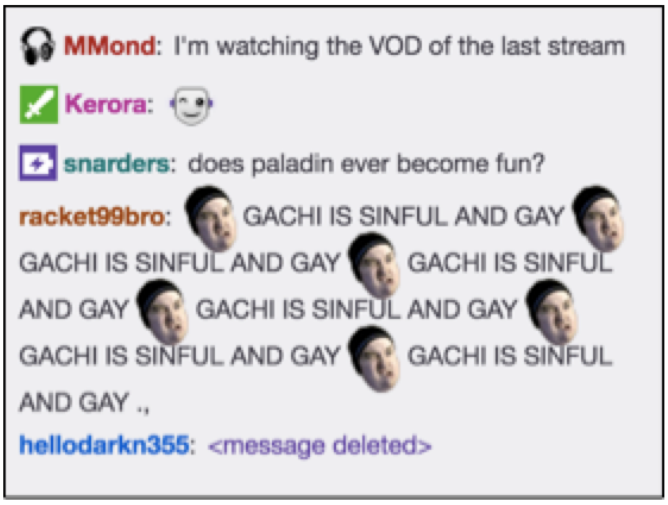

Figure 2: User Types and Behaviors

teristics based on the text of the message. These tags included flags for each of the three types of behavior described above: spam, questions, and smiles. Each message was also tagged with the number of messages being sent in the channel per second at the time it was sent, which was used as a measure of how many users were actively chatting in the channel at any given time. Table 2 shows the variables used in this study.

Figure 2 shows examples of each type of user and each type of content. The first message in Figure 2 has none of the content types we flagged, and was posted by a subscriber. Each subscriber icon is unique to the channel to which it is attached; in this case, the subscriber icon for a music channel is a set of headphones. The second message is a smile posted by a moderator, indicated by a green sword icon. The third message is a question posted by a Twitch Turbo user, as indicated by a purple battery icon. The last two messages were posted by regular users. The first of these has spam with many capital letters and emotes, and the second is marked as deleted because of a ban.

In order to analyze the impact of specific events on the messages that followed them, messages were grouped into clusters of twenty-one for all analyses except the analysis of moderation mode. These clusters contained ten messages to establish the state of the channel prior to the an event of interest, one "event" message with prop-

erties that served as independent variables, and ten messages following the "event" with characteristics that served as dependent variables. We included all messages as events that had enough messages sent before them and after them to use in analysis. In the corpus of approximately 21 million messages, we analyzed approximately one million event messages that had ten messages prior and ten subsequent to each. In the moderation mode analysis we analyzed groups of forty messages: twenty before an event and twenty after an event. In this case the event in question was a change in chat moderation mode that occurred in between the twentieth and twenty-first messages.

The "Rate" variable used in this analysis, which notes how many messages per second were being sent in a given chatroom at the time of the event, is used here as a proxy for size of channel. We used rate as measure of size instead of number of viewers because different channels may have higher or lower rates of participation, and total viewer numbers can be skewed by tools like viewbots, which are used to artificially inflate the number of viewers in a channel. In contrast, the rate of messages being sent at a given time models the user's perception of the size and level of activity of the crowd. Across all analyses, a higher rate of messages being sent was associated with more spam, fewer questions, and fewer smiles.

These analyses all use interrupted time-series models as described in Nagin [24]. Shadish, Cook, & Campbell [31] describe two major challenges in proving causality. First, when it is argued that an event A caused an event B, it must be shown that the reverse explanation could not be true (i.e. B did not cause A). The interrupted time-series approach resolves this problem by analyzing events that occur in fixed temporal sequences; B could not have caused A because B occurred after A.

The second challenge to proving causality is disproving the possibility of alternative explanations based on outside events. In this case, the most plausible outside explanation is that behavior displayed on the video stream, which was not captured in this dataset, explains the relationship between event messages and subsequent chat behavior. This possibility will be explored in more depth in the final section of this paper, but it does not conflict with the implications of this work.

Variable Name	Description	Mean
PriorState	Number of messages of a given type (spam, questions, smiles) in the past ten messages. Sum of Boolean for each of the ten prior messages. Centered at mean = 0.	Spam: 1.48 Questions: 0.47 Smiles: 0.09
Event	Whether a given message contains spam/question/smile. 1 = YES, 0 = NO	Spam: 0.148 Questions: 0.047 Smiles: 0.009
Rate	Messages per second being sent in the chat at the time of the event. Centered at mean = 0.	4.9
isMod	Whether a given message was sent by a moderator. 1 = YES, 0 = NO	0.083
isSub	Whether a given message was sent by a subscriber. 1 = YES, 0 = NO	0.235
isTurbo	Whether a given message was sent by a Turbo-user. 1 = YES, 0 = NO	0.036
isBanned	Whether a given message was banned. 1 = YES, 0 = NO	0.021
R9k	Whether a given message was sent in R9kbeta mode. 1 = YES, 0 = NO	0.154
Slow	The number of seconds of slow mode currently on. If slow mode is off, equals zero.	13.82
Sub	Whether a given message was sent in subscribers-only mode. 1 = YES, 0 = NO	0.028

Table 2: Variables, Descriptions and Means

All data in this study was explored in aggregate, and no individual message text was analyzed by the researchers beyond scripted tagging of messages according to the three content types. Messages in the dataset were tied to users only through pseudonymous usernames, which were eliminated from the final dataset used for analysis. No experiments were performed, and no users encountered a different experience on the website than they normally would have.

4.4 Analysis 1: Effects of Imitation

Before attempting to understand the impact of key individuals and moderation techniques on imitation, we need to first demonstrate that imitation occurs in this setting. Thus to test H1, we analyzed the impact of "event" messages on subsequent content.

Table 3 shows the regression coefficients for a linear regression on percentage of a given type of behavior in the ten messages following an event as a function of properties of this event, prior state, and rate of messages at the time of the event. Note that the PriorState and Rate variables were centered to have a mean of zero, which causes the intercept to be equal to the percentage of messages of the given type in the next ten messages in a channel with average characteristics when the "event" message has none of the possible characteristics. This linear regression followed the format:

$$P' = Intercept + PriorState + Event^{5} + Rate$$

Using the coefficients listed in Table 3, the percentage of messages containing spam in the next ten messages following an event is 13.7% + 5.7% times the number of spam messages above the average amount of spam in the prior ten messages plus 6.0% if the event message was spam, plus 0.5% per additional message per second being sent at the time of the event above the average rate.

[5]"Event" is a binary variable noting whether the event message contained the type of content being analyzed in the given column. For example, event would equal one in the left column if the message contained spam, one in the middle column if the message contained a question, and one in the right column if the message contained a smile. In this case, if the event message contained spam, an additional 6% of the subsequent ten messages contained spam.

	Dependent Variables:					
	Percentage Spam		Percentage Questions		Percentage Smiles	
Independent Variables:	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept	13.7%***	0.01%	4.7%***	0.00%	1.0%***	0.00%
PriorState	5.7%***	0.01%	2.3%***	0.01%	2.0%***	0.01%
Event[1]	6.0%***	0.03%	2.6%***	0.03%	2.2%***	0.03%
Rate	0.5%***	0.00%	-0.3%***	0.00%	-0.7%***	0.00%
N=2032436	$R^2 = 0.44$		$R^2 = 0.11$		$R^2 = 0.06$	

* p < 0.05 ** p < 0.01 *** p < 0.001

Table 3: Impact of a given type of "event" message on the subsequent ten messages

Behavior	% Increase after Event
Spam	43.8%
Question	55.3%
Smile	220.0%

Table 4: Percentage Increase after Event of Same Type

In each case, regardless of whether an event message contained spam, questions, or smiles, the following ten messages had a significantly higher proportion of messages of that same type than if the event message had not been of that type, after controlling for rate of messages in the chat and prior quantity of each type of behavior. Of these, smiles were imitated significantly more than either other type. This confirms H1, as these behaviors display imitative properties. In addition, more rare posts were imitated at a higher rate.

Table 4 shows the percentage increase in each type of behavior following the event. For example, when an event message contained spam, we observed 43.8% more messages containing spam over the next ten messages. Of the three types of messages, smiles were most susceptible to these imitation effects and spam was the least susceptible.

4.4.1 Discussion

The above analyses provide strong evidence for the presence of imitative effects in Twitch chatrooms, supporting H1. New users may be introduced to various types

of behaviors by observing others engaging in them. Existing users may be reminded about particular behaviors or encouraged to engage in them when other users do so. These effects can be seen as a combination of the conception effect [35], where users are reminded of the possibility of engaging in a particular behavior, and broad imitation effects. In this case the most common behavior showed the smallest increase in percentage as a result of an imitation effect. This may be the result of small impact of a conception effect; because spam is so common, users do not need to be reminded that they can post spam themselves. Since smiles are quite rare, observing a smile does remind users of a mode of interaction that they might not otherwise have considered.

Of the above behaviors, questions are perhaps the least intuitively likely to be imitated, as each question expects a response that is not in the form of a question. One plausible scenario for imitation within the category of questions is that when a user receives a response to a question from the streamer or a moderator, other users ask questions because they believe they are likely to receive answers.

4.5 Analysis 2: Impact of User-Type

In our second analysis, we further explored the imitation observed in Analysis 1 to test H2, that individuals with greater status and authority would be imitated more frequently.

We used four categories of users as examples of different levels of status and authority within a specific ingroup. Moderators show both status and authority; subscribers show status; regular users show neither commitment nor authority; and Twitch Turbo users show status, but not within the particular ingroup. In this sense, Twitch Turbo users can be compared to Wikipedia editors who are not attached to the specific project at hand, but still have some status [36]; Turbo users are not attached to the specific channel in which they are chatting, but they have a mark of status.

In this analysis, as in the first analysis, groups of twenty-one messages were analyzed: ten messages establishing the prior state of the channel at the time of the

	Dependent Variables					
	Percentage Spam		Percentage Questions		Percentage Smiles	
Independent Variables	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept	14.0%***	0.01%	4.8%***	0.01%	0.9%***	0.00%
PriorState	5.6%***	0.06%	2.3%***	0.01%	2.0%***	0.01%
$Event^2$	5.4%***	0.03%	2.5%***	0.03%	2.1%***	0.03%
isMod	-0.7%***	0.04%	-0.04%*	0.02%	0.2%***	0.01%
isSub	-0.9%***	0.02%	-0.4%***	0.01%	0.04%***	0.01%
isTurbo	-0.4%***	0.05%	1.9%***	0.03%	0.03%**	0.01%
Rate	0.5%***	0.00%	-0.3%***	0.00%	-0.1%***	0.00%
Event*isMod	4.8%***	0.13%	1.2%***	0.12%	0.7%***	0.08%
Event*isSub	2.0%***	0.07%	0.07%	0.07%	-0.01%	0.06%
Event*isTurbo	-2.9%***	0.18%	-1.4%***	0.17%	-0.63%***	0.12%
N=2032436	$R^2 = 0.44$		$R^2 = 0.11$		$R^2 = 0.06$	

* p < 0.05 ** p < 0.01 *** p < 0.001

Table 5: Impact of a message posted by a given type of user on subsequent messages of the same type

event, one event message with characteristics that were treated as independent variables, and ten subsequent messages with characteristics that served as dependent variables.

Table 5 shows the regression coefficients for a linear regression on percentage of the given type of behavior in the ten messages following an event as a function of this event, prior state, rate of messages at the time of the event, type of user, and interaction between type of user and event. This linear regression followed the format:

$$P' = Intercept + PriorState + Event + isMod + isSub + isTurbo + Rate + Event*isMod + Event*isSub + Event*isTurbo$$

The results show that certain types of users had more impact than others. Status effects were persistent across all three categories of behavior, except in the case of Turbo users who had status but not within the ingroup. Table 6 shows the impact of a given type of user posting a given type of content on percentage of that content over the subsequent ten messages. For example, when a moderator posted a message containing spam, 67.8% more messages with spam were posted in the next ten messages, while if a user with no authority posted a message with spam the number of

Behavior	User Type	% Increase after Event
Spam	Mod	67.8%
	Sub	46.4%
	Turbo	15.0%
	Regular User	38.6%
Question	Mod	76.3%
	Sub	45.2%
	Turbo	62.5%
	Regular User	52.1%
Smile	Mod	333.3%
	Sub	236.7%
	Turbo	166.7%
	Regular User	233.3%

Table 6: Percentage Increase after Event of Same Type Posted by Given User Type

messages with spam in the next ten messages increased by 38.6%.

4.5.1 Discussion

Overall, different types of users were imitated at substantially different rates across the three categories of behaviors, supporting H2. This suggests that imitation and conformity to authority effects exist in this context.

However, Twitch Turbo users, who paid monthly to have privileges across the site, were imitated less than regular users across all three categories of behavior. This could be explained by a perception of Turbo users as outsiders; because they openly display commitment to the whole site as opposed to one channel, they may be viewed as lacking commitment to the channel in which they are chatting. Zhu, Kraut, and Kittur [36] found that, among Wikipedia workers collaborating on projects, editors who identified with the projects acted similarly to prototypical group members, but editors who didn't identify with the group did not act similarly to prototypical group members at all despite their apparent status. This suggests that unless high status individuals also show commitment to the local community, members of the community will not imitate them. Alternatively, Twitch Turbo users may feel that their Turbo subscription gives them some authority that other users do not think that they have. The data supports this second explanation; overall, users with only the Twitch Turbo tag were banned at approximately five times the rate of subscribers, who were almost never banned. While Twitch Turbo users may have some additional status as a result

of their badge, regular users can plainly see that this badge does not afford them special treatment. For the purposes of influence, they are nearly as much "outsiders" as regular users.

Moderators were imitated significantly more than non-moderators when posting spam, questions, and smiles. Subscribers followed the same pattern as moderators, though they were significantly less influential in all three cases, and not statistically significantly influential on smiles or questions.

4.6 Analysis 3: Effects of deterrence

In the third portion of our analysis we explored the impact of moderation behaviors on subsequent message characteristics. Literature on deterrence suggests a number of ways that direct and indirect experiences with punishment might affect future behavior [22, 34].

4.6.1 Impact of chat moderation modes on behavior

The first part of this analysis looks at behavior before and after a particular chat moderation mode was enabled in a chat in order to test H3, that chat moderation modes will deter spam but will not affect other types of behaviors.

By typing a command, channel owners and moderators can enable chat modes in a channel that restrict the types of messages that can be sent. Three chat modes were explored for this analysis: subscribers only mode, where only users who have subscribed to the channel may chat; slow mode, where users may only post messages every N seconds, where N is set by the moderator who enabled the mode; and R9K beta mode, where users who post messages of 9 characters or more are only permitted to post messages that have not previously been posted.[6]

These analyses were performed on groups of forty messages, where a change in channel mode occurred between the twentieth and twenty-first message in the set. The

[6]This last mode is named after an experimental bot in a webcomic forum where it was first explored `https://web.archive.org/web/20190618113538/https://blog.xkcd.com/2008/01/14/robot9000-and-xkcd-signal-attacking-noise-in-chat/`.

Category	Mode type	Change in next 20	Percent Change	P(Δ= 0)
Spam	Slow	-0.52	-14.3%	0.048*
	Sub	-0.95	-22.7%	0.003**
	r9k	-0.46	-14.7%	0.033*
Questions	Slow	0.16	14.6%	0.277
	Sub	-0.17	-18.9%	0.206
	r9k	0.27	31.3%	0.243
Smiles	Slow	0.08	38.4%	0.383
	Sub	0.10	66.4%	0.306
	r9k	0.08	66.1%	0.414

* p < 0.05 ** p < 0.01 *** p < 0.001

Table 7: Percentage Change in Behaviors Following Implementation of a Chat Mode

first twenty messages were compared to the second twenty messages to get an idea of channel behavior before and after the switch. The use of forty messages instead of twenty here is a response to the significantly smaller sample size; chat mode changes were relatively rare in this dataset. Slow mode, subscribers-only mode, and R9K beta mode were enabled 168 times, 667 times, and 97 times respectively in scenarios where there were twenty messages before and after the change during a single stream.

Overall, enabling each of these modes led to substantial decreases in spam in subsequent messages, but had no significant effect on subsequent behaviors of the other types. This supports H3.

Table 7 shows the impact of enabling chat modes on subsequent behavior. For example, after enabling subscribers-only mode there were 0.95 fewer messages containing spam in the next twenty messages, a decrease of 22.7%.

4.6.2 Impact of bans on subsequent imitation

To test H4, that bans would succeed in discouraging multiple types of behaviors we looked at the impact of banning a particular type of behavior on the frequency of that type of behavior in the next ten messages. We analyzed approximately two million groups of twenty-one messages. As in previous analyses, these groups consisted of ten messages prior to the event to establish a baseline for behavior at the time of the

	Dependent Variables					
	Percentage Spam		Percentage Questions		Percentage Smiles	
Independent Variables	**Coefficient**	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
Intercept	13.7%***	0.01%	4.7%***	0.01%	1.0%***	0.00%
PriorState	5.6%***	0.06%	2.3%***	0.07%	2.0%***	0.01%
Event[3]	6.4%***	0.03%	2.6%***	0.02%	2.2%***	0.03%
isBanned	0.1%	0.09%	0.6%***	0.04%	0.01%***	0.02%
Rate	0.5%***	0.00%	-0.3%***	0.00%	-0.1%	0.00%
Event*isBanned	-4.7%***	0.13%	-1.6%***	0.02%	-2.3%**	0.20%
N=2032436	$R^2 = 0.44$		$R^2 = 0.11$		$R^2 = 0.06$	

* p < 0.05 ** p < 0.01 *** p < 0.001

Table 8: Impact of banning a particular type of message on subsequent messages

event, an event message with characteristics from which independent variables were drawn, and ten subsequent messages with characteristics that served as dependent variables. In approximately 2.3% of these cases, the event message was banned.

In nearly all cases, these data represent the impact of generalized deterrence âĂŞ the indirect experience with punishment. Because users who were banned are prevented from posting for period of time, other users typically posted the subsequent messages. These results represent the impact of observing another user being banned for a particular type of behavior.

Table 8 shows the regression coefficients for a linear regression on percentage of a given type of behavior in the ten messages following an event message as a function of event type, prior state, rate of messages at the time of the event, whether a message was banned, and interaction between ban and type of message. This linear regression followed the format:

$$P' = Intercept + PriorState + Event + isBanned + Rate + Event*isBanned$$

Table 9 shows the impact of bans on frequency of a particular behavior. Overall, banning any type of behavior had a significant negative impact on the frequency that behavior appeared in subsequent messages, confirming H4. For example, when a message containing a smile was banned, the percentage of smiles in the next ten

Behavior	Response	% Increase after Event
Spam	Banned	13.1%
	Not Banned	46.7%
Question	Banned	34.0%
	Not Banned	55.3%
Smile	Banned	-10.0%
	Not Banned	220.0%

Table 9: Percentage Increase after Event of Same Type, Banned vs Not Banned

messages decreased by 10.0%, but increased by 220.0% when the message was not banned.

Because bans did not necessarily occur directly after a message was posted, it is reasonable to question whether the amount of time between message posting and ban had an effect on subsequent messages. However, in our analyses we found no statistically significant impact. Given that the majority of bans in this dataset occurred within one second of message posting, chat moderation bots probably administered most of the bans automatically. In particular, in larger channels with higher rates of posting there was a higher proportion of rapid ban responses indicating more bot involvement. This suggests that broadcasters and moderators make use of available tools to administer bans before many more posts have been made. In slow chats, human moderators can accomplish this, while faster chats require moderation bots. This effect may be more important than the absolute amount of time between posting and ban.

While this analysis shows significant generalized deterrence effects, these effects are not as large as imitation effects for the three types of content; even though a message was banned and thus disappeared from the screen, its brief presence reminded users that this behavior was one possibility in which they could engage, suggesting the presence of the conception effect [35].

4.6.3 Discussion

The analyses presented above show significant but not uniform deterrent effects from bans, and a significant decrease in spam behavior during chat moderation modes.

The deterrence effects of seeing another person being banned for engaging in a

particular type of behavior is in line with literature on generalized deterrence [13, 34]. While deterrence literature suggests that the effect of general deterrence decreases as social distance between observer and criminal increases, such distinctions are less relevant on Twitch. Users directly observe the punishment of users who are mostly socially indistinguishable from them, and they are not at all affected by bans that they don't observe in other networks (i.e. other channels).

In this case, channel chat moderation modes had limited effects. While they were successful in deterring one type of behavior, spam, their lack of flexibility prevented them from being applied differentially to discourage or encourage other types of behavior. In this regard, regular bans were more flexible, but none of the moderation tools were effective in encouraging positive behaviors (i.e., smiles).

4.7 Analysis 4: Duration of Impact

In the final part of this study we examined how long imitation effects last. We calculated correlations between counts of each type of behavior in two blocks of ten consecutive messages with a varying number of messages between them, using these correlations across channels to measure imitation effects. We looked at decay in the effects by increasing the delay between the set of messages that were the source of the effect and the set of messages that displayed imitation. Table 10 shows correlations between counts of behavior in blocks of ten messages separated by zero, ten, twenty, thirty, forty, and fifty messages. For example, the correlation between count of spam messages in one block of ten messages and count of spam messages in the next block of ten messages was 0.29, but the correlation between count of spam messages in one block of ten messages and count of spam messages in another block of ten messages 50 messages later was 0.15. For all three behaviors, there was a nonzero correlation between behavior in the initial block and behavior in the next block, but this decreased as the next block moved further away. This shows that these behaviors have significant short-term effects but that these effects decrease steadily over time.

This matches the results discussed earlier both in Analysis 1 and in Bakshy et al.

	Interval Size					
Behavior	**0**	**10**	**20**	**30**	**40**	**50**
Spam	0.29	0.23	0.20	0.17	0.17	0.15
Questions	0.13	0.11	0.09	0.08	0.07	0.07
Smiles	0.10	0.07	0.06	0.05	0.05	0.03

Table 10: Correlation between count of messages of each type in one group of ten messages and the count of messages of same type in the next group of ten at increasing intervals between groups

[8]. Users were influenced to post content that they saw other users post, and the more other users posted the content the more likely they were to follow suit. However, this effect was limited to a relatively brief period of time after the initial posting.

These correlations were calculated for each of the 400 channels in our sample and for each of the three behaviors described above: "spam", "questions", and "smiles". Approximately 2 million comparisons were included in each of these analyses.

Taken together, these results suggest that a small cluster of messages of a particular type may have some short-term impact, but is unlikely to change the culture of the channel. Such a cultural shift might plausibly occur as a result of a larger number of clusters of particular types of behavior, especially if such behaviors were encouraged by users with authority or left unbanned.

4.8 Conclusions

The studies described above show clear patterns of imitation and deterrence in Twitch chats. When a user posted a message with a particular type of behavior, subsequent messages were substantially more likely to contain that behavior. This is consistent with much research on imitation.

Beyond this general imitation effect, our research shows that distinct types of users were more influential on certain types of behaviors. Overall, users with more authority had more influence on most types of behavior. However, Twitch Turbo users had very little influence or even negative influence, which may reflect their outsider status.

Finally, users were responsive to deterrent measures. When a particular type of behavior was banned, the frequency of that behavior in the next group of messages decreased. When chat moderation modes were enabled, most types of behavior were unaffected by all three modes but spam decreased in all cases.

These findings have substantial practical implications. First, in designing future moderation tools, it may be useful to consider the possibility of encouraging desirable behaviors in combination with the standard method of punishing undesirable ones. While Analysis 4 shows minimal long-term impact of imitation effects on culture, it may be possible to curate a culture with repeated interventions over time. By actively banning undesired content and encouraging users to behave positively toward each other, site managers may be able to create a community that is resistant to the impact of undesirable behaviors from newcomers or outsiders who aim to disrupt. Conversely, once a channel has developed a norm of undesirable behaviors, it may be more difficult to stop these behaviors from spreading.

These findings also point to several areas for future research. Visible examples of good and bad behavior have substantial impact [21]. Future studies could experiment with different approaches to making good and bad behaviors more visible. In addition, while the proactive tools explored here (i.e., the chat modes in Twitch) are relatively inflexible, alternative tools could be tested that warned users in advance if their message was likely to be banned based on its content. Moreover, new tools that offer broadcasters more flexibility in determining which types of behavior to discourage may be more effective in regulating culture. While there may be a role for machine learning-driven approaches to regulating specific types of behavior, such approaches are inherently risky in that they may drive users away because of unclear or inconsistent standards for appropriate behavior or may encourage some users to be increasingly creative in their attempts to be offensive in order to circumvent automated bans. A promising direction for future exploration is use of social pressure to enforce standards; allowing established users in the community more visibility and a more active role in discouraging unwanted behavior, even if not directly through bans, can turn a chatroom full of viewers into a group of allies.

As discussed above, one of the primary difficulties in establishing causality is elimination of outside explanations. Due to limitations of the IRC medium, this analysis does not control for behaviors exhibited through video or audio on the stream itself. As such, one alternative explanation for the imitation effects observed is that users were more likely to engage in certain behaviors when they received a cue from the streamer that such behaviors were acceptable, and that more established users were more sensitive to these cues. Broadcasters who are more calm and collected may recruit and retain viewers who are more likely to behave calmly [15]. While we cannot separate this influence out in our analysis, the implications of this explanation are mostly the same; in both cases, streamers can significantly affect chat behaviors both through encouraging examples of acceptable behavior and by using tools to deter unacceptable behavior, regardless of whether the reaction comes from a moderator in the chat or the streamer.

Our research suggests that existing moderation tools can be effective and, by extension, that moderators and broadcasters have some ability to shape the type of chat environment that they want, though they may still be vulnerable to persistent campaigns of targeted harassment. The combination of moderation tools described in this paper can help channels of almost any size; smaller channels will see more impact from bans because the next ten messages will last longer. Larger channels may benefit more from a combination of chat modes and bans.

More broadly, these findings have implications for how moderation should be explored both in research and in practice. Moderation can be viewed not only as a reaction to specific events but also a method for preventing the spread of unwanted behavior and development of undesirable norms for what conduct is acceptable. The development of behavioral standards through display of positive behaviors is possible both independent from and in combination with moderation tools.

Bibliography

[1] RL Akers, WF Skinner, MD Krohn, and RM Lauer. Recent trends in teenage tobacco use-findings from a 5-year longitudinal-study. *Sociology and Social Research*, 71(2):110–114, 1987.

[2] Ronald L Akers and John K Cochran. Adolescent marijuana use: A test of three theories of deviant behavior. *Deviant Behavior*, 6(4):323–346, 1985.

[3] Ronald L Akers and Anthony J La Greca. Alcohol use among the elderly: Social learning, community context, and life events. *Society, culture, and drinking patterns reexamined*, 1:242, 1991.

[4] Ronald L Akers and Gang Lee. A longitudinal test of social learning theory: Adolescent smoking. *Journal of drug issues*, 26(2):317–343, 1996.

[5] AnyKey. Workshop #3 White Paper: Barriers to inclusion and retention: The role of community management and moderation. Technical report, 2016.

[6] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639, 2011.

[7] Solomon E Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9):1, 1956.

[8] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

[9] Joseph B Bayer, Nicole B Ellison, Sarita Y Schoenebeck, and Emily B Falk. Sharing the small moments: ephemeral social interaction on snapchat. *Information, Communication & Society*, 19(7):956–977, 2016.

[10] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[11] Robert B Cialdini. *Influence: Science and practice*, volume 4. Pearson education Boston, 2009.

[12] Sauvik Das, Adam DI Kramer, Laura A Dabbish, and Jason I Hong. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1416–1426. ACM, 2015.

[13] Joan Feigenbaum, James A Hendler, Aaron D Jaggard, Daniel J Weitzner, and Rebecca N Wright. Accountability and deterrence in online life. In *Proceedings of the 3rd International Web Science Conference*, page 7. ACM, 2011.

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

[15] William A. Hamilton, Oliver Garretson, and Andruid Kerne. Streaming on twitch: Fostering participatory communities of play within live mixed media. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1315–1324, New York, NY, USA, 2014. ACM.

[16] George E Higgins, Brian D Fell, and Abby L Wilson. Digital piracy: Assessing the contributions of an integrated self-control theory and social learning theory using structural equation modeling. *Criminal Justice Studies*, 19(1):3–22, 2006.

[17] George E Higgins, Abby L Wilson, and Brian D Fell. An application of deterrence theory to software piracy. *Journal of Criminal Justice and Popular Culture*, 12(3):166–184, 2005.

[18] Charles K Hofling, Eveline Brotzman, Sarah Dalrymple, Nancy Graves, and Chester M Pierce. An experimental study in nurse-physician relationships. *The Journal of nervous and mental disease*, 143(2):171–180, 1966.

[19] Michael A Hogg. A social identity theory of leadership. *Personality and social psychology review*, 5(3):184–200, 2001.

[20] Aniket Kittur, Bryan Pendleton, and Robert E Kraut. Herding the cats: the influence of groups in coordinating peer production. In *Proceedings of the 5th international Symposium on Wikis and Open Collaboration*, page 7. ACM, 2009.

[21] Robert E Kraut and Paul Resnick. *Building Successful Online Communities: Evidence-based Social Design*. MIT Press, 2012.

[22] Xigen Li and Nico Nergadze. Deterrence effect of four legal and extralegal factors on online copyright infringement. *Journal of Computer-Mediated Communication*, 14(2):307–327, 2009.

[23] Stanley Milgram and Christian Gudehus. Obedience to authority, 1978.

[24] Daniel S Nagin. Criminal deterrence research at the outset of the twenty-first century. *Crime and justice*, 23:1–42, 1998.

[25] Jessica M Nolan, P Wesley Schultz, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. Normative social influence is underdetected. *Personality and social psychology bulletin*, 34(7):913–923, 2008.

[26] Douglas P. Peters and Stephen J. Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187âĂŞ195, 1982.

[27] David P Phillips. The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American Sociological Review*, pages 340–354, 1974.

[28] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

[29] Bridie Scott-Parker, Melissa K Hyde, Barry Watson, and Mark J King. Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding? *Accident Analysis & Prevention*, 50:242–250, 2013.

[30] Joseph Seering, Robert E Kraut, and Laura Dabbish. Shaping Pro and Anti - Social Behavior on Twitch Through Moderation and Example - Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017.

[31] William R Shadish, Thomas D Cook, and Donald T Campbell. Experimental and quasi-experimental designs for generalized causal inference. 2002.

[32] Muzafer Sherif. The psychology of social norms. 1936.

[33] Mark C Stafford and Mark Warr. A reconceptualization of general and specific deterrence. *Journal of research in crime and delinquency*, 30(2):123–135, 1993.

[34] Michael Tonry. Learning from the limitations of deterrence research. *Crime and Justice*, 37(1):279–311, 2008.

[35] Ladd Wheeler. Toward a theory of behavioral contagion. *Psychological Review*, 73(2):179, 1966.

[36] Haiyi Zhu, Robert Kraut, and Aniket Kittur. Organizing without formal organization: group identification, goal setting and social modeling in directing online production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 935–944. ACM, 2012.

Chapter 5

Proposed work

5.1 Background

As social media has expanded globally, content moderation has become increasingly important to the health of platforms and public discourse. In the previous chapters I have outlined the current state of moderation research, the different perspectives and their implications. While all of these perspectives are valuable, I have elected to focus on the community moderation perspective because I think it has the most potential to help address the largest outstanding problems in moderation. As noted earlier, major platforms such as Twitter, Instagram, and YouTube use centralized moderation strategies that combine human flagging with algorithmic detection methods to identify potentially problematic content. This content is either handled automatically or passed on to human reviewers, who are typically employees or contractors for the company. However, this model has not yet been scaled to a degree where it can effectively moderate the massive volumes of content generated on these platforms; algorithms have proven fairly effective in detecting spam and pornographic content, but they cannot address the social nuances associated with more complex misbehavior like hate speech or harassment or bullying.

The inability of existing content moderation systems to handle complex social issues at the necessary scale is a major social problem. Examples of the impact of problematic content that has slipped through the cracks are trivial to find; moreover,

platforms have struggled to find a consistent process for handling important cases that are brought to their attention. This is not a failure of ingenuity on their part – making decisions about nuanced social situations replete with cultural and historical context is impossible to do at the necessary speed. Facebook simply cannot employ enough people to thoughtfully mediate all of the world's conflicts. From my work, I see the potential for volunteer community moderation to play a larger role in managing online interactions. The moderators I interviewed handle complex social situations on a daily or weekly basis, and they are as well-situated as anyone to understand the context. The primary advantage of increased reliance on volunteer user moderation from the platforms' perspectives is that it can allow these platforms to focus on moderating *communities* rather than *users*; the volunteer moderators will address the vast majority of incidents on their own, and the company only needs to take action when moderators in a particular community cannot (or choose not to) handle a situation that has gotten out of control.

The application of volunteer community moderation systems to a broader variety of contexts is not a simple task, nor is it one that would immediately solve all problems with moderation online. However, based on my current work I believe that a shift toward greater user participation in making moderation decisions would switch the problems mentioned earlier in this work with a somewhat more manageable set that can be addressed through design and development of new systems and tools. Rather than identifying hate speech or harassment in single pieces of content or a small history of interactions, platforms can focus on identifying communities that regularly allow or host this content; speaking purely from a technical standpoint, it is much easier to algorithmically flag potentially-problematic communities at scale than to classify individual pieces of content as problematic simply because each instance of the former case has orders of magnitude more data to analyze. Moreover, rather than focusing on trying to discourage users from posting problematic *content* at a massive scale, platforms can work to help moderators respond to problematic *interactions* on a much smaller scale. As I note in chapter three of this proposal, there are many situations and levels of interaction where moderators could be better supported or

given better tools.

Though my prior work provides a foundation, major questions remain. From a full-system perspective, volunteer moderatorsâĂŹ current relationships with platform administrators need to be understood in order to consider how such relationships might be redesigned in the future. On a more sociotechnical level, the ideal division of labor between moderators and algorithmically-driven tools has not yet been clearly defined, and the factors that would shape this division are not yet clearly outlined. Finally, though the work described above identified some preliminary social and technical approaches to improving moderation, there remains much space for development of deeper and more nuanced solutions. In my dissertation I will address each of these three areas in order to develop a comprehensive review of the potential for future application of the community-driven moderation model. I will approach these questions from three directions.

5.2 Part One: Longitudinal investigation of moderators' practices

Chapter three presented a study based off of interviews with 56 moderators from Reddit, Facebook Groups, and Twitch. I propose to build on this work by re-interviewing these moderators approximately two years after their initial interviews to better understand how their experiences have been shaped over time as the platforms have evolved. There are several reasons I have chosen to focus on a longitudinal study. First, none of the very few studies of moderators conducted in the past twenty years has followed the trajectory of these moderators. While popular discussion of moderation has talked about moderator "burnout" because of emotional labor, abuse, or overwork,[1] no research has yet provided concrete generalizable evidence about why moderators stop moderating. I will seek to re-interview these previous moderators regardless of whether or not they are continuing to moderate, and I will ask them

[1] see e.g., https://web.archive.org/web/20190723144905/https://www.engadget.com/2018/08/31/reddit-moderators-speak-out/

questions about reasons they have for moderating more, fewer, or different groups than they previously moderated. If the practices of volunteer user moderation are to be supported or even expanded, supporting longer retention of moderators will be an important goal.

Second, while some prior work has looked at how users adopt various features (e.g., [2]), and a variety of work has looked at the impact of implementing new technology in workplaces, no work has yet explored how users adopt tools that are designed to help them manage casual social spaces. Twitch, Reddit, and Facebook Groups have all added a significant number of new features that impact moderation in the past two years. For example, Twitch has added several features including "AutoMod", an algorithm for automatically flagging and holding potentially problematic content that is ostensibly driven by some sort of AI, though the specific details are not transparent to users. Reddit has undergone the largest UI overhaul in its history, where workflows have changed and several new tools have been added. Facebook Groups have also seen changes, including streamlining the process for making violators aware of what rules they have broken and allowing moderators to keep better logs of users' past offenses.

Another major change in the way moderators use moderation tools has resulted from these platforms' gradual shift over time from desktop to mobile engagement. Reddit launched an official mobile app for the first time in its history in 2016 (having previously allowed third-party apps to provide users with mobile access to the platform), and utilization of this app has grown steadily since. Facebook has also increased its focus over time on mobile use of the platform. To date, no research has focused on volunteer moderation research performed on mobile apps, but it is plausible that this difference could change moderators' workflows significantly.

One of the most important "feature" changes, however, has been the expansion of Discord into the space of online communities. Discord is a platform that is functionally similar to Slack, with channel-based chatroom style conversations, but with a greater emphasis on voice communication. Rather than conducting a new study specifically on Discord, it is useful to study Discord in conjunction with these existing

platforms because of the way Discord has grown; much of Discord's early growth resulted from creation of Discord groups that were associated specifically with existing Twitch channels or subreddits. As such, I expect a significant fraction of the moderators I previously interviewed on these two platforms to now have some experience with Discord, whether as a user or a moderator in spaces connected to their pre-existing communities. This will allow me to study cases of inter-platform moderation, a case which has not yet been examined in depth in any research.

A final reason for conducting longitudinal studies is the changes in policies community guidelines and policies that have been made on these platforms over the past two years. About a year after I completed my interviews of Twitch moderators, Twitch announced two major policy changes. The first described in more depth what types of sexual conduct were permitted on the platform, and in the second Twitch reserved the right to remove users for their behavior on other platforms[2], e.g., Twitter.[3] Reddit has also become significantly more active over the past two years in banning subreddits that host extreme content. For example, subreddits advocating harassment and rape of women (/r/incels) and violence toward political liberals in America (/r/Physical_Removal) were banned in late 2017. The /r/gore and /r/WatchPeopleDie subreddits were banned in March of 2019 in response to alleged glorification of the Christchurch mosque shootings in New Zealand. Most controversially, the main subreddit for Donald Trump supporters on Reddit, /r/The_Donald, was quarantined in June 2019 for "threats of violence against police and public officials". While there has been much focus in the literature about the importance of site policies for conduct [3, 6], and Matias has argued that moderators' maintenance of their relationships with platform administrators are an important part of their job, I will seek to understand how these policy chances have actually impacted the moderation practices of moderators of average subreddits.

[2]To my knowledge, no other major platform has yet adopted this same stance.

[3]Over the past few days, Twitch has attracted controversy by banning users who posted videos on Twitter joking about animal abuse.

5.2.1 The division of labor between humans and algorithms (and platforms)

One of the major questions in designing tools to support volunteer moderators is what work can be done automatically and what work is best left to moderators and their capacity to make judgments based on context. In the work I presented in chapter three, I argued that tools should "support, rather than supplant, the judgment of users" [19, p. 1419]. Similarly, in subsequent work Jhaver et al. argued that automated tools need to be designed with careful attention to how they will impact the workload of moderators [4, p. 31:26]. In these follow-up interviews, I will ask moderators to discuss each of the moderation-related tools and features that have been added to the platform. I will ask them to answer:

1. To what extent they have utilized each of these new features
2. Why they have chosen not to use certain features
3. In what cases they use each feature, and how this process is different from how they previously handled these cases
4. To what extent prior problems in their communities have been resolved by these tools (drawing from problems they had highlighted in their previous interview)
5. What fraction of their current time they spend moderating through tools vs through social engagement

I will also further probe their responses to these questions based on whether they moderate primarily on the desktop client, the mobile app, or a mix of the two.

Though the above questions have focused on division of labor between humans and algorithms, it is also important to consider what authority (and thus workload) platforms delegate to volunteer moderators and what authority they retain for themselves. In the work I presented in chapter three, I noted that my interviewees identified a "clear division in labor between community moderators and platforms". They felt that "it is their job (and often their right) to manage their communities and that

platform employees should only intervene if things go very wrong" [19, p. 1426–1427]. This, however, is not how platforms operate. For example, prior to quarantining /r/The_Donald,[4] Reddit representatives repeatedly told users during Q&A sessions that they had been working with the subreddit's moderators to address problematic behaviors within the subreddit, but these efforts were not ultimately successful and the inability for these two groups to work together successfully led to the quarantining of the subreddit.

It is unrealistic for moderators to expect to have the final say over moderation in their communities; platforms are owned by companies, and these companies have universally reserved the right to remove users or content at their discretion. Volunteer moderators are therefore situated lower on the moderation hierarchy than platform administrators, but the relationship between these two groups has not yet been defined with any formality outside of extreme cases like the above. If volunteer community moderation is to be integrated more deeply and perhaps more formally into the platforms' moderation strategies, potential relationships, responsibilities, and avenues for communication should be explored. In order to begin to understand these possibilities, I will also ask moderators the following questions:

1. Has [the platform] ever intervened within your community to moderate a user or content that has been posted? [If so, describe...]

2. Have you ever been in a situation where you needed help from the platform in order to address a moderation issue?

   (a) In what ways did you communicate the situation and your needs?

   (b) In what was was the response from the platform useful or not useful?

3. What factors do you think platform administrators should take into account when deciding when to intervene in your community?

With this work I aim to develop a clearer model of how specific tools interact with social processes in order to define moderators' workflows, as well as how effec-

[4]This is the previously-mentioned primary subreddit for Donald Trump supporters

tive relationships might be constructed between volunteer moderators and platform administrators. Future tools that are a net drain on moderators' time (e.g., when moderators on Reddit have to correct many false positives flagged by Automoderator [4]) will be less likely to be adopted. Perhaps more importantly, tools that shift moderators' workflows away from engaging directly with users and toward engaging indirectly with users through tools could harm communities' long-term social development in ways that might not be obvious to designers or even the moderators themselves. Moreover, the relationships between volunteer moderators and platform administrators will play a significant role in the effectiveness of any formal attempt to more fully integrate volunteer moderation into the broader systems of platform moderation.[5]

5.2.2 Metaphors for moderation

A variety of literature has demonstrated the importance of mental models and metaphors in how people understand and make sense of systems and complex tasks [7]. For example, Lareau [8] found that the metaphors parents used to describe how they saw the task of child-raising were directly connected with the life outcomes of their children. Metaphors are also core to various interaction design processes [13]; Lockton's recent *New Metaphors* method uses metaphors to inspire new and unusual combinations of ideas [9]. In my first round of interviews, I heard users describe their work in a variety of metaphorical terms, even without my prompting. One moderator said that he saw himself as a gardener, removing "weeds" and encouraging "flowers". Another said that sometimes he felt like his job was to be a piñata. A third compared his team to lawmakers in government but noted that they didn't have the power of a "line-item veto". These metaphors all reveal depth in how moderators see their work that they may not otherwise be able to verbalize, and the consideration of these metaphors can help give moderators heuristics to know how to act in unfamiliar situations.

[5]Volunteer moderators are by definition volunteers and not employees, but in aggregate they perform a huge volume of work that brings value to these community-based platforms. One of the reasons that these relationships may not yet have been developed formally is that this could highlight the tension inherent in a for-profit company's heavy reliance on volunteers.

One of the major processes that I identified as part of my "Moderator Engagement Model of Community Development" [19] was the process of learning how to be a moderator. As I noted in chapter three, there is almost no formal onboarding process for moderators on any of these platforms, so moderators are left to act according to their personal values and their perceptions of the group. This leads to a fairly steep learning curve in figuring out when to intervene and how to intervene in order to best address incidents in the short term and help the community develop in the long term. As the final portion of my follow-up interviews, I will directly ask moderators about the metaphors that they would use to describe their moderation styles and philosophies. I will also ask moderators to choose a metaphor that best fits their mental model of how the platform makes moderation decisions:

1. If you had to use a metaphor to describe your philosophy for moderation and how you see your role as a moderator, what would it be?
    (a) How specifically does this metaphor describe the ways you moderate?
    (b) Has the role you just described changed over time as you've grown as a moderator, or would you say that this metaphor has been consistent since you began as a moderator?
2. If you had to pick a metaphor to describe how [the platform] moderates, what would it be?

I will formally group the metaphors moderators provide for their own philosophies and roles and connect them with specific practices, leading to a formal categorization scheme for metaphors that moderators use to describe their work.[6] This scheme has two potential uses. First, in designing tools to support these moderators, it will help platforms to have a high-level sense of how different types of moderators frame their behaviors and understand their roles. Tools can be tailored to specific approaches rather than attempting to design one-size-fits-all tools. Second, these metaphors can

[6] I have already begun grouping and classifying the metaphors that moderators used in the first round of interviews from two years ago.

be used as part of a more formal onboarding process for moderators; helping them to consider ahead of time what roles they might take can help give them a head start in learning how to deal with difficult situations.

5.3 Part Two: Analysis of language in rules and norm-setting

In the second portion of my proposed work, I will move toward more direct quantitative analyses of social strategies for improving online behaviors. Prior work has demonstrated that the presence of clear, visible guidelines for how to behave can measurably reduce problematic behaviors [5, 11]. In line with this work, community-based platforms have developed features to highlight rules. The default UI for subreddits has included for many years a box where moderators can list rules; users can see the subreddit's rules in the sidebar on every post and on each subreddit's landing page. Twitch added a feature within the past two years that allows channels to write a list of rules that pop up before a new user types their first message in the chat. Facebook Groups also recently streamlined the process for helping moderators let violators know which group rules they have broken. However, no research has yet explored how to write rules effectively. In order to address this question, I will perform two quantitative analyses based on Reddit and one set of experiments.

5.3.1 Understanding rebukes

In the first analysis I will seek to understand what language regular users currently use to rebuke each other in online discussions. While other work has looked at persuasive language, e.g., Yang and Kraut's work on Kiva lending teams [21] and Tan et al.'s work on the /r/ChangeMyView subreddit [20], I choose to look at language focused specifically on addressing perceived transgressions. I will work from a definition of "rebuke" offered by Radzik:

"A rebuke is a pointed expression of disapprovalâĂŤusually but not neces-

> *sarily moral disapprovalâĂŤ-that is addressed to a perceived transgressor [...] One way or another, it communicates to the perceived wrongdoer the message that the action was wrong (morally or otherwise) and that she was responsible for it."* [12, p. 644]

Rather than beginning by defining rebukable content and examining responses, I will begin by collecting a large volume of Reddit comments and labeling rebukes independent of the content that they are responding to. I do this in order to avoid beginning this process with a narrow definition of what types of behavior can be rebuked; based on my work and other prior work, I know that a wide variety of behaviors can reasonably be rebuked, from hate speech to harassment to advertisements for t-shirts and pictures of boxes, and effective rules should be able to deal with more than just the obvious cases. Based on these comments, I will develop a classifier that can identify with reasonable accuracy whether a comment is a rebuke, and I will collect and label a large corpus of comments.[7]

Within this corpus, I will identify threads that contain both a rebuke and a response to this rebuke with the goal of identifying the impact of different approaches to rebuking. I will also measure other relevant factors, e.g., the relative score (i.e., upvotes and downvotes) of each of these comments, which may indicate the local community's perspective on the disagreement, and the overall score of the post, which may indicate the visibility that the post got outside its normal audience due to popularity. Ultimately, the goal of this work is to identify linguistic characteristics of effective rebukes with the hypothesis that these factors may help either the development of better rules or more effective phrasings for moderators to use when explaining to users what they have done wrong.

5.3.2 Linguistic factors in rule impact

Next, I will perform a similar analysis looking at the impact of rules on "rebukable" behaviors. I will identify 200 subreddits and match them on measures of similarity

[7] I have begun the early phases of collecting comments and building a classifier as advised by Diyi Yang.

using the approach used in Chandresekharan et al.'s quantitative analysis of norms on Reddit [1]. I will collect three months of data from these subreddits. I will then manually collect the language used in the rules for each of these subreddits, and will perform analyses to determine how variations in rules for similar communities are associated with different frequencies of rebukable behaviors. In order to do this, I will create a general classification schema for the most common types of rules, e.g., "Rules about seriousness of comments" or "Rules about abusive behaviors". I then will use the above pairwise matching approach to compare the effectiveness of wordings of different subreddits' rules focusing on these same categories, i.e., what is the optimal way to phrase a rule about abusive behavior?

This is not intended to be a causal analysis, but it aims to gather evidence about relationships between language and behavior that can be used in subsequent causal analyses.

5.3.3 Applying linguistic principles to improve rule development

Building from the findings from the above two studies, I will run an experiment using the fake political blog post experimental structure that I previously used in my work with using CAPTCHAs to shift user behavior (Seering et al., 2019 [14]). This structure exposes users to a fake political blog post with fake comments listed below. Users are asked to contribute their own comment to the thread. See Figure 5.3.3 for more detail.

I will design multiple sets of rules with different linguistic properties that I expect to lead to different behavioral outcomes, and show each set of rules to participants before they write their comment response. I will analyze comments based on the same techniques previously used in my aforementioned work.[8] If the analyses from the first two parts here shows that different language is differently effective across

[8]Note that I have attempted a first trial of this experiment without support from the previous analyses I describe here to build from, but it was not successful in identifying any interesting differences between different sets of rules. My hope is that evidence from the first two parts of this section will help me create more meaningful conditions.

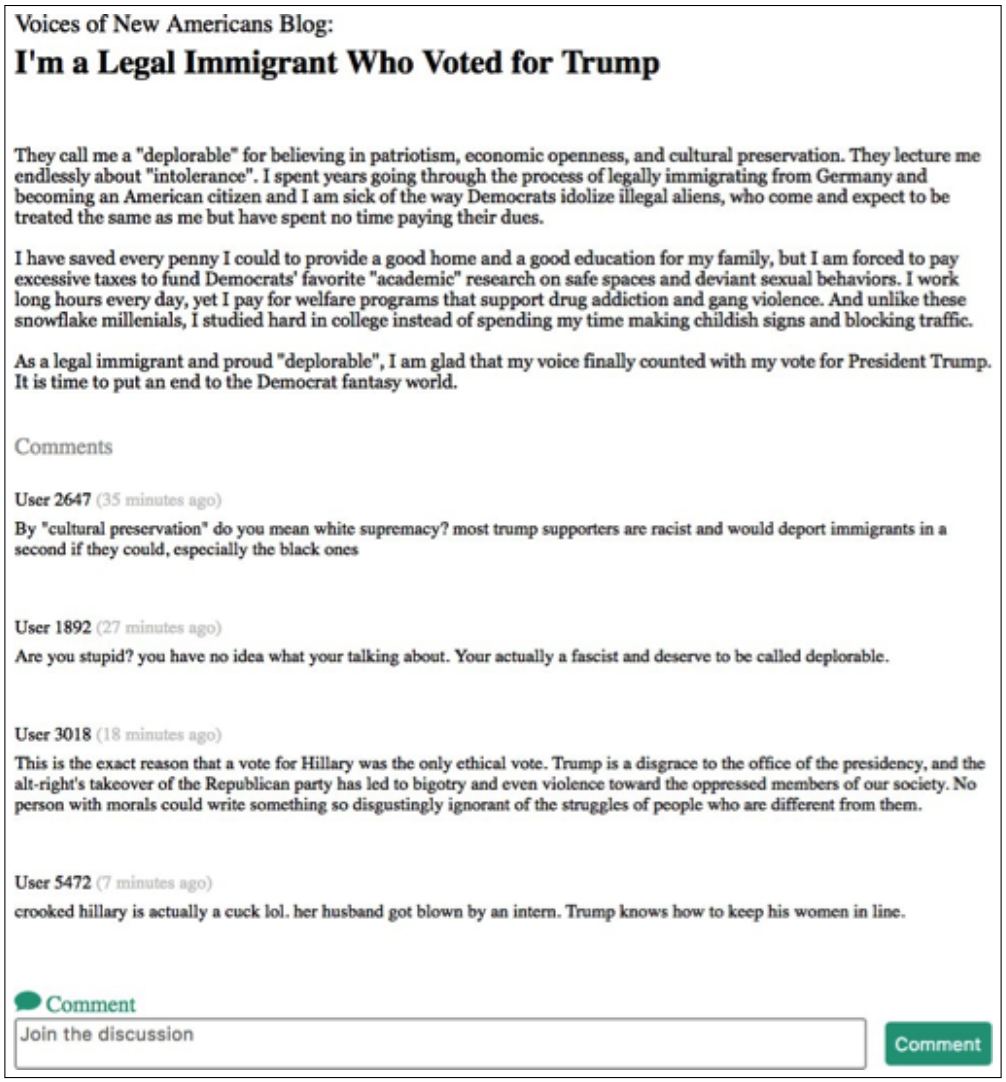

Figure 5-1: Final version of comment thread page

categories of rules, e.g., that certain language is more effective in rules about abuse but less effective in rules about off-topic comments, I will test multiple categories of rules in this experiment.

5.4 Part Three: Proactive norm-setting and identity-building through conversational agents

The previous sections of this chapter have focused analyses of users' and moderators' behaviors aimed at understanding processes or improving social approaches to mod-

eration. This thesis proposal as a whole has so far been generally skeptical about the value of jumping toward technical solutions for these major social problems; my discussion in chapter two of the value and efficacy of algorithms for detecting problematic behaviors is the main example of this skepticism. I have also noted the relative rarity of proactive approaches to moderation, with exceptions being my work on imitation [16] and the CAPTCHA work described above [14]. In this section I propose to modify widely-used conversational agents ("chatbots") to act in ways that proactively encourage positive behaviors.

Though the majority of this proposal has built from the prior work I have done in the space of moderation, I have also worked on three projects that were spun off from my analysis of social behaviors on Twitch. In exploring the data I collected from Twitch chatrooms for my work on imitation, I found a significant volume of messages sent by bots. On Twitch [15], Reddit [10], and now Discord, many if not a majority of moderators use bots to perform moderation actions, post information, engage users, or run simple games. I found in my analyses of Twitch bots' actions that bots were actually the most "active" participants in channels of any size and topic by more than an order of magnitude above the average for any other type of user [15, p. 157:12]. However, the behaviors they displayed were notably simplistic; likely due to the interface that less-technically-proficient moderators use to manage the most popular bots, messages were extremely similar across channels often with only one or two words changed.

In the second piece of my work on bots, I first performed a literature review of chatbot work in the ACM space to identify research that has studied or developed sophisticated social behaviors for chatbots based in communities [17]. I found that the vast majority (more than 90%) of past chatbot research in this space has focused on *dyadic* interactions in chatbots – conversations taking place between one user and one bot. Even when bots were placed within community-style platforms, e.g., Slack, their messages were typically either posted in response to a command sent by a user or were directed at a single user. Very little work has focused on chatbots "hanging out" in communities or interjecting organically into group discussions, and no work

has explored this in a technically-sophisticated way. This is not a trivial challenge by any means; turn-taking in conversation is a very complex social phenomenon, where people take cues from social status, body language, conversational topic, vocal tones, spatial and cultural context, and a variety of other factors. I concluded this piece of research by proposing social roles that chatbots might take in communities in the future.

So, to summarize the above: chatbots are widely used by moderators in major online communities, but their behaviors are socially unsophisticated and rely on the same general conventions that chatbots have used in one-on-one conversations for more than fifty years. In my previous work I have identified how these bots are used and how they fit into moderators' workflows [15, 19], and I have proposed designs for social roles they might take in the future [17]. I have also explored behavioral approaches that they may drawn from, e.g., imitation effects [16] and general social identity principles [18], and in the work that I have proposed above I aim to identify linguistic patterns that may also be of use.

Over the past six months I have built a functional prototype of a chatbot that can be used as a technical base for these goals, and I am currently finishing a three-week pilot test in a real Twitch community with approximately twenty active members. This bot, currently nicknamed "baby_bot", is designed to learn language from its community over a relatively long period of time, growing from a bot "baby" to a toddler, and eventually through its teenage years. I have used a mix of rule-based methods and ML-based methods (currently Markov chains with topic seeding) to generate text. I have chosen the child-raising theme for a number of reasons. First, I hypothesize that the process of raising a "child" will help a community bond and consider their values as a community. Second, this framing allows for a dynamic example of how training data can and should be used in developing an AI. While other previous examples of ML-driven chatbots deployed in the wild have shown how such systems can be abused,[9] I believe that established communities can be much more thoughtful and civil in raising a chatbot child; as with a human child, a chatbot

[9] https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course

is likely to turn out better if raised (trained) by a community than if let loose in the Twittersphere. Third, creating a chatbot that begins as a child allows its designers to make use of the conceit of ignorance. Just as we wouldn't expect a human child to be born with full adult conversational skills, users will interpret a baby chatbot's initial messages through this lens, making them more tolerant when it says things that are ignorant of social conventions. This allows us as designers to test behaviors that diverge from how both bots and other users might normally behave.

Given that most of the core development work has already been done for this chatbot, I propose as my final piece of work in my dissertation to use it as a testing platform for the findings from the other projects I have proposed here. I will use it to test different linguistic patterns, social behaviors, and structures for different relationships with moderators, all in live deployment on Twitch. Though the specific metrics I will use to analyze the bot's success will depend on the results of the above studies, I will likely perform both qualitative and quantitative analyses.

5.5 Conclusions

By the end of this work, I aim to be able to determine to what extent and in what contexts the expansion of community-driven moderation can be a feasible solution to growing problems in online content moderation. In addition to the above studies, I hope to conclude with a broad discussion of the advantages and disadvantages that community-driven approaches to moderation have in different situations from the perspectives of different stakeholders. From this perspective, I will comment on future directions in research and development that might strengthen the position of community-driven approaches to moderation in relation to the alternatives.

Bibliography

[1] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's

hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32, 2018.

[2] Sauvik Das, Adam DI Kramer, Laura A Dabbish, and Jason I Hong. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 1416–1426. ACM, 2015.

[3] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.

[4] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2019.

[5] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design*, 2012.

[6] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 2018.

[7] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

[8] Annette Lareau. *Unequal childhoods: Class, race, and family life*. University of California Press, 2011.

[9] Dan Lockton. New metaphors, 2018.

[10] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. Could you define that in bot terms?: Requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI*

Conference on Human Factors in Computing Systems, pages 3488–3500. ACM, 2017.

[11] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 2019.

[12] linda Radzik. Moral rebukes and social avoidance. *The Journal of Value Inquiry*, 48(4):643–661, 2014.

[13] Dan Saffer. The role of metaphor in interaction design. *Information Architecture Summit*, 6, 2005.

[14] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–14, Glasgow, Scotland, UK, 2019. ACM.

[15] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The social roles of bots: Evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):157, 2018.

[16] Joseph Seering, Robert E Kraut, and Laura Dabbish. Shaping Pro and Anti - Social Behavior on Twitch Through Moderation and Example - Setting. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, 2017.

[17] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 450. ACM, 2019.

[18] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. Applications of social identity theory to research and design in computer-supported cooperative

work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):201, 2018.

[19] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society*, 2019.

[20] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee, 2016.

[21] Diyi Yang and Robert E Kraut. Persuading teammates to give: Systematic versus heuristic cues for soliciting loans. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):114, 2017.

Appendix A

Volunteer Community Moderator Interview Protocol

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

A.0.1 Introduction

1. How long have you been active on [the platform]?

2. How long have you been a moderator for [specific community]?

3. Do you moderate any other communities? For this interview, please focus your answers on [community].

A.0.2 Primary Questions

1. How would you describe the culture of this community?

2. How did you become a moderator in [community]? [If they started the group: How did you select moderators for your community? General characteristics, needs, experience?]

(a) Did you know other moderators in the group before you became a moderator?

(b) Did you have prior experience as a moderator?

(c) Were you active in the community before becoming a moderator? If so, in what ways?

(d) Did you provide support to the community such as design, technical, financial, or other ways?

(e) Did you volunteer or were you asked to become a moderator?

3. How did you learn about how to be a group moderator?

   (a) Were you ever formally trained by someone on how to be a moderator in the group?

   (b) Did you receive any instructions?

   (c) Did any other moderators give you guidelines or advice?

   (d) Were you given examples of scenarios that might come up and instructions on how to handle them?

   (e) Did you get a chance to practice in any way before starting to moderate the community?

   (f) Did you learn anything about how to be a moderator from [site]'s tutorials or explanations of moderation resources?

4. Do members of your team have specific roles? [If single moderator, or they say they all moderators do the same things: Can you tell me about the different types of things you do in managing this community?]

5. In what types of situations with the group do you have to step in as a moderator?

6. What types of violations have you spent the most time dealing with in the past week or so?

7. *Double check with them* - "It sounds like you do x, y, and z. Is that everything, or is there anything else?"

8. What are some technical tools you use to make your job easier?

    (a) Do you use any sort of screening to filter new members?

    (b) Do you find these tools to be sufficient?

9. In the past month, what sort of things have you spent the most time discussing with other moderators of the group?

    (a) What platform(s) do you use to discuss these things?

10. Do regular members of your community ever help with moderation?

    (a) Do they ever criticize people who break rules or explain to them how to behave in the group?

    (b) Do they ever report content that they think violates the rules?

11. Do you ever warn users before bans or post removals?

    (a) When are these warnings customized, and when are they templates?

12. Do you ever post or send explanations of the rules after removing a post?

    (a) When are these explanations customized, and when are they templates?

13. Can you give me an example from the past week or two of how an offender has reacted to being punished?

    (a) Is this a typical reaction?

14. How have the rules in your community changed over time?

    (a) What is the process for changing rules like?

15. Have you ever interacted with [platform] employees regarding your group? If so, about what?

(a) How frequently do you think platform admins review your community for compliance to site-wide content policies?

16. Can you describe to me a significant or memorable moderation experience that you've had in this group? It can be positive or negative.

A.0.3 Conclusion

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

Appendix B

Volunteer Community Moderator Interviewee Characteristics

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

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

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

Appendix C

Volunteer Community Moderator Interview Code counts

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

Figure C-1: Steps and variants in Being and Becoming a Moderator process [$\kappa = 0.89$]. Code counts by Twitch (T), Reddit (R), and Facebook (F)

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

Figure C-2: Steps and variants in Moderation Tasks, Actions, and Responses process [$\kappa = 0.70$]. Code counts by Twitch (T), Reddit (R), and Facebook (F)

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

Figure C-3: Steps and variants in Rules and Community Development process [$\kappa = 0.85$]. Code counts by Twitch (T), Reddit (R), and Facebook (F)