Supporting Volunteer Moderation Practices in Online Communities

Joseph Seering

CMU-HCII-20-107
September 2020

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
jseering@andrew.cmu.edu

Thesis Committee:

Geoff Kaufman (Chair)
Human-Computer Interaction Institute
Carnegie Mellon University

Jason Hong
Human-Computer Interaction Institute
Carnegie Mellon University

Bob Kraut
Human-Computer Interaction Institute
Carnegie Mellon University

Michael Bernstein
Department of Computer Science
Stanford University

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

KEYWORDS
Content moderation, platforms, social computing, computer-supported cooperative work, social media, volunteer moderators, cooperative responsibility, Twitch, Reddit, Facebook, AI-mediated communication, social identity theory, rebukes, interpersonal moderation, governance, communities, online communities, self-governance, self-moderation, literature review, platforms and policies, commercial content moderation, human-centered design, metaphors, computational social science, thematic analysis, interviews, digital labor, hate speech, harassment, social networks.

Abstract

In this dissertation, I explore multiple levels of the content moderation ecosystem with a focus on platforms that rely extensively on volunteer user labor. These platforms, like Reddit, Twitch, and Facebook Groups, expect users to moderate their own communities, but reserve the right to intervene when communities or content therein violates sitewide standards for behavior.

This thesis contains three parts. I begin with a high-level exploration of how platforms do and do not engage with volunteer community moderators. I build on the framework of cooperative responsibility to analyze the different ways platforms and users have found common ground on values, roles, and spaces for deliberation. Next, I focus in depth on the philosophies and mental models of the volunteer moderators, analyzing the metaphors they used both explicitly and implicitly to describe the work they do. Finally, I dive into the specifics of interpersonal language use in moderation, looking at how both interpersonal "rebukes" impact subsequent comment threads on Reddit and how changes of rules in communities on Reddit impact subsequent behavior of community members. For each of these linguistic pieces, I present results from a related experiment, in which I used a custom-built comment forum to test the impact of simulated "rebukes" and rules.

This work shows the nuance of several core processes in user-driven moderation, ranging from the very high level organizational interactions to very low level linguistic features of user comments, and I argue that more attention toward understanding these processes in Computer-Supported Cooperative Work and related fields is needed.

Contents

1

Introduction: The Content Moderation Ecosystem

1.1 Prelude[1]

In May of 1978, the "CommuniTree #1" online Bulletin Board System (BBS) launched in the San Francisco Bay area [6, pp. 88–92], [7]. 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 social spaces; while the first set of these virtual bulletin boards, developed in the mid-to-late 1970s, only displayed messages either in alphabetical order or in the order messages were posted [7], CommuniTree #1's tree-style design allowed for conversations to move fluidly in multiple directions. The CommuniTree #1 platform and its subsequent iterations were infused with its creators' philosophy of the grand power of social technology – the first discussion thread (called a "conference") opened with the bold statement, "We are as gods and might as well get good at it". The participants (mostly academics, researchers, and computing hobbyists) 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" [6, p. 90][2]. In 1982, Apple entered into an

[1] A modified version of this chapter combined with parts of the second and sixth chapters has been accepted to CSCW 2020 with Joseph Seering as solo author.

[2] Reflecting the "Digital Utopianism" [8] of this era of technologists, a technical manual written for CommuniTree by Dean Gengle was dedicated to "R. Buckminster 'Bucky' Fuller / The first global shaman of our species." [3, p. iii]

agreement with the United States government to provide schools with Apple computers as a substitute for paying taxes, which caused a huge influx of teenage, mostly male users into virtual spaces previously reserved for the intellectual elite. Upon discovering CommuniTree, these students filled the board's allotted disk space with "every word they could think of that meant shitting or fucking" [7], an onslaught for which existing users were completely unprepared. CommuniTree had been launched with minimal moderation tools; an "anti-censorship" philosophy was written directly into its code, with features that prevented system operators from proactively filtering messages as they came in, made it difficult to remove messages once they were entered, and granted any user access to commands that controlled the host computer, so 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 almost entirely 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 may be one of the earliest major failures of online moderation documented in research literature, and, at least in hindsight, was a major blow to the dream that the internet could function simply as a 'marketplace of ideas' where better perspectives would naturally rise to the top. Today, popular media is full of examples of problematic behaviors, including extensive harassment leading users to delete their accounts,[3] or adopt defensive behaviors [2, 9]. A 2017 PEW study found that 4 in 10 Americans had personally experienced online harassment [5]. Major platforms have closed their comment sections or forums because of an inability to maintain positive and productive conversations, including NPR,[4] Popular Science,[5] and IMDB.[6] Despite tremendous growth in the adoption of online social systems, now used by a 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.

[3] E.g., https://web.archive.org/web/20190722033005/https://www.buzzfeednews.com/article/krishrach/people-are-upset-after-kelly-marie-tran-deleted-her on Instagram

[4] https://web.archive.org/web/20190723073441/https://www.npr.org/sections/publiceditor/2016/08/17/489516952/npr-website-to-get-rid-of-comments

[5] https://web.archive.org/web/20190712210415/https://www.popsci.com/science/article/2013-09/why-were-shutting-our-comments/

[6] https://www.theverge.com/2017/2/3/14501390/imdb-closing-user-forums-comments

1.2 Three Levels of Content Moderation

Content moderation is a process that occurs on many different levels, includes many different groups and actors, and handles tremendously important social problems. In this dissertation, I provide examples of ways to evaluate content moderation at various levels, and discuss how these levels are interrelated. I argue that, while much of the public focus has been on the platform-wide level, it is imperative for the future of content moderation that the full ecosystem be understood.

Content moderation can be understood as occurring on "micro", "macro", and "meso" levels, terminology elaborated upon in depth for the moderation context by Chandrasekharan et al. [1]. Taking a slightly different approach from Chandrasekharan et al., who focused primarily on norms, I define the macro, meso, and micro levels of moderation as that which are done by platforms and governments, community moderators, and individual users respectively.

In this dissertation I examine each of these levels and, to some extent, their interplay. I argue that not only is it important to understand each, but that it is impossible to truly understand the content moderation ecosystem without looking at the overlap between the levels. Successful future interventions in content moderation will be designed based on consideration of the motivations, behaviors, processes, and values of agents at all three levels, with earnest attempts to map the interplay between each of these factors across each of the levels.

Large platforms cannot realistically parse every piece of content posted to their sites in the depth needed to incorporate an understanding of local and cultural context into moderation decisions,[7] and it is unlikely that this capacity will be developed in the foreseeable future. Given this, the present moment is an appropriate time to consider the future of moderation in online social spaces from a broader perspective, looking at the different systems and models that have developed across various types of platforms at each level of the moderation ecosystem.

[7] I refer to moderation that incorporates this understanding as "context-sensitive" moderation.

1.3 Dissertation Outline

This dissertation contains six chapters, which includes the present chapter, four chapters of work I describe below, and a concluding chapter that discusses the implications of this work more broadly.

The second chapter of this dissertation contains a literature review comparing literature in an "institutional agents" perspective – the perspective that focuses on the processes of platforms, governments, and NGOs – with a "communities perspective", which focuses on the moderation processes within user-run communities. I argue that the latter has been studied extensively, but that its lessons have not yet reached the literature in the former perspective.

The third chapter of this dissertation is an evaluation of major platforms' strategies for engaging users on topics surrounding content moderation, using Helberger, Pierson, and Poell's concept of *cooperative responsibility* [4]. I use data both from interviews with moderators on three platforms and from analysis of documents published by these platforms, that progress has been made toward a balanced division of roles and responsibilities but that there is potential for improved processes in this regard.

The fourth chapter explores the cognitive models and metaphors that volunteer community moderators hold, drawing data from 79 interviews with volunteer moderators from Facebook, Reddit, and Twitch. I identify more than 20 metaphors commonly but variably used across these platforms. I argue that these metaphors can be generative in further understanding moderators' work practices and generating new perspectives for research, but also for inspiring new designs for moderation tools that support moderators whose work has previously gone largely unnoticed. I present two potential designs for new types of moderation tools based off of these metaphors.

The fifth chapter provides the results of four studies looking at the use of language in norm-setting and rule-setting actions. I first analyze the use of rebukes on Reddit, identifying comments where one user rebukes another user for an expressed opinion or behavior, and analyzing the outcomes resulting from different approaches to rebuking. Second, I present results from a controlled experiment looking at how differently worded rebukes impact sub-

sequent behavior of onlookers. Third, I analyze the processes of changing rules on Reddit in order to try to understand how rule changes impact subsequent behaviors. Finally, I present results from another controlled experiment where the text of rules presented to participants was varied in an attempt to shift their behaviors.

In the concluding chapter, I discuss potential future work in this space. I provide six directions in which moderation research could expand, particularly in the space of studying moderation within online communities, and I comment broadly on the importance of this work in shaping the future of social behaviors online.

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. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):32:1–32:25, November 2018.

[2] 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.

[3] Dean Gengle. *Communitree*. The CommuniTree Group, San Francisco, CA, USA, first edition, 1981.

[4] Natali Helberger, Jo Pierson, and Thomas Poell. Governing online platforms: From contested to cooperative responsibility. *Information Society*, 2018.

[5] Pew Research Center. Online harassment 2017. Report, Pew Research Center, Washington, D.C., 2017, July.

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

[7] Allucquère Rosanne Stone. What vampires know: Transsubjection and transgender in cyberspace. Talk Given at the "In Control: Mensch-Interface-Maschine" Conference in Graz, Austria, 1993.

[8] Fred Turner. *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. University of Chicago Press, Chicago, IL, USA, 2010.

[9] 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*, CSCW '17, pages 1231–1245, New York, NY, USA, 2017. ACM.

2

Foundations: Two Perspectives in Moderation Research

2.1 Research in Content Moderation[1]

Though moderation has recently drawn interest from researchers working from many different perspectives, the school of academic thought that has shaped public discourse most in the past several years operates from a *platforms and policies* perspective focused on *platforms* and *governments*, as well as some influential *non-governmental organizations*, as the dominant drivers of moderation rather than users. Though some of the major researchers working from this perspective have critiqued this model and proposed ways it might be improved, most of these researchers at least implicitly presume that moderation in the future will be driven by powerful centralized agents like governments and platforms. Academics writing from this platforms and policies perspective have proposed legal and structural fixes which, despite their imperfections, fit a clear and evolving narrative. The relative lack of influence of researchers working from *user* and *community-based* perspectives on the public discourse may be caused in part by their reluctance to propose broader solutions that extrapolate beyond their empirical findings. Thus, the primary goal of this chapter is to provide a clear

[1] A modified version of this chapter combined with parts of the first and sixth chapters has been accepted to CSCW 2020 with Joseph Seering as solo author.

outline of the state of community moderation research in the context of other domains.

In the following sections, I contrast the *platforms and policies* and *communities* perspectives within moderation research, writing as a researcher who identifies with the latter approach. The former body of research presumes a largely top-down model of moderation as the default, asking questions like 'How can Facebook make transparent content moderation decisions?' [19, 81] and 'How might laws be written to shift platforms' content moderation processes in more productive directions?' [13]. The communities perspective focuses instead on the self-moderating social structures, i.e., communities, through which people interact online, with the goal of understanding how these structures function and how they might be improved. This perspective asks questions like "What are the different roles volunteer moderators play in online communities?" [57, 86] and "How do these moderators make rules?" [75]. Note that these perspectives are not mutually exclusive, and a valuable goal for future work would be to integrate the two. However, in order to attempt to integrate these perspectives we must first understand the value and contribution of both.

The ways I define "moderation" research and the perspectives within it are only one way of organizing this space.[2] Though the researchers whose work I discuss here might reasonably take issue with how their work is categorized, I am not claiming that this is in any way *the* definitive approach to categorizing work in this field. Other ways of organizing the research to date would almost certainly yield other sets of insights. I have chosen to bound and classify prior work in a way that helps draw a roadmap for where the field of community moderation research can go and what work is needed to address some of the major challenges to a more community-driven model that can take a greater role in the content-moderation ecosystem.

[2]The "moderation" research domain is very broad and does not have sharply defined edges, so choices must be made regarding where to bound a literature review. For example, I have chosen *not* to cover disinformation behaviors or those that fall within the domain of cybersecurity, though my recommendations have implications for those domains.

2.2 Two Perspectives in Moderation Research

This section details the two perspectives in depth: the **platforms and policies** perspective and the **communities** perspective. I describe overlap between these perspectives as appropriate.

2.2.1 The Platforms and Policies Perspective

Research in the *platforms and policies* perspective focuses on content moderation from the perspectives of online social platforms, governments, and sometimes nonprofits, frequently under the implicit assumption that these centralized agents will be the primary shaping force in the future of the governance of online speech.

Organizational and Legal Perspectives:

In his 2018 book *Custodians of the Internet*, Communications & Media scholar Tarleton Gillespie argues that social media platforms are caretakers of the online world [32]. They are custodians both in the sense that they must keep these platforms *clean* and in that they have *custody* of modern discourse. He discusses the broader challenges platforms face in balancing their general philosophies with ethics, technical feasibility, and legal requirements, and how these challenges manifest in how they write *community guidelines*, how they decide *what content to moderate*, and how they *structure their organizations*. These topics are all major focus areas within the platforms and policies perspective in moderation research.

Gillespie's argument in *Custodians* built significantly from his earlier publication, "The Politics of 'Platforms"' [31], which described platforms as inherently political entities that have attempted to maintain an image of neutrality. The late 2000s to early 2010s saw a major transition from "online communities" to "platforms" and "social networks", changing both the structures of social relations online and the language used to describe them. Gillespie analyzed the underlying meaning of the word "platform", which had become the dominant

label for describing Facebook, YouTube, and other rapidly-growing sites by the time the paper was published in 2010, arguing 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 visible modern work taking a platforms and policies perspective, other scholars have made important arguments from this same general perspective. For example, 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 [47]. Klonick discusses legal regulation of content moderation in the United States including the history of Section 230 of the Communications Act, which specifies protections for platforms engaged in moderating users' speech. Though originally intended to provide "a limited safe harbor from liability for online providers engaged in self-regulation" [14, p. 455], i.e., platforms that engaged in "good faith" content moderation, it has been interpreted by courts in ways that give platforms extensive leeway in what and how they moderate.[3] The European Union's General Data Protection Regulation (GDPR), which was passed twenty years after Section 230 and went into effect in mid-2018, focuses on a different but related area of platform responsibility – namely, the protection of users' data [1, 21]. Despite its different focus, it is clearly a stark ideological contrast to Section 230. Whereas Section 230 grants platforms extensive leeway, GDPR requires platforms to meet strict standards for user privacy.

Section 230 has been the focus of other legal scholars' work in topics related to moder-

[3]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

ation; Danielle Keats Citron's *Hate Crimes in Cyberspace* discusses the role of Section 230 in protecting platforms that host "revenge porn" [12]. Both Citron and Mary Anne Franks have written in depth about potential approaches to regulating revenge porn and how these approaches would interact with Section 230 [24, pp. 1282–1291].[4] Citron and Franks, along with Benjamin Wittes, are among a small group of legal scholars who have been openly critical of the leeway granted to platforms under Section 230 [13, 14] [25, pp. 161–181]. Citron and Wittes argue that "The Sky Will Not Fall" if interpretive or even carefully-written federal statutory changes are made to Section 230 [13, p. 411].

While the above authors differ somewhat in how they approach platforms' roles, all focus on the substantial social, technical, and political power that modern platforms wield and the political challenges they face. The need for platforms to navigate political responses to their policies is a challenge that has grown significantly in importance since 2017. Due to the role of platforms in hosting political ads and general political speech, politicians have a vested interest in platforms' rules permitting types of content that benefit them. This has caused political conflicts over fact-checking policies and general policing of truth, as well as the types of political ads that platforms agree to host. Recent work by Douek [19] analyzes a response by Facebook to these challenges, which is the creation of an oversight board that may have some say over difficult moderation decisions. Though the specific form that the Board will take and the impact of its decisions remain to be seen, Douek makes the general point that this oversight board cannot be an appeals process (due to the massive volume of potential appeals) or an "ultimate arbiter of free speech norms" [19, pp. 6–7], but rather should aim to reduce blind spots that Facebook has in its rule-making processes and serve as an independent forum for discussion.

Another prominent recent perspective drawing from legal traditions comes from David Kaye in his 2019 book *Speech Police: The Global Struggle to Govern the Internet*. Kaye

[4]The work of Citron and Franks has been very influential; per https://web.archive.org/web/20191106224932/https://www.businessinsider.com/map-states-where-revenge-porn-banned-2019-10/, 46 states plus Washington D.C. had laws against revenge porn as of the end of October 2019, up from two states in 2013.

makes a number of proposals for improving content moderation [43, pp. 112–126], many of which match calls from Gillespie, Klonick, Douek, and others for increased transparency and accountability. He also calls for increased decentralization in platforms' decision-making processes, though by this he means greater involvement by "local" stakeholders in a *geographic* sense rather than in the virtual community sense discussed in the following section. He argues that "companies should make human rights law the explicit standard underlying their content moderation and write that into their rules" [43, p. 119]. Platforms have struggled to find a universal set of principles to shape and justify their content moderation decisions; to this end, Kaye suggests that human rights law is an appropriate place to start. However, this approach, combined with the decentralization of decision-making that Kaye proposes, will require the creation of a complex and multi-layered system of global governance built with platforms at its center. This new speech-governance body could plausibly be one of the most ambitious international governance projects in human history, with stakeholders from around the world lobbying to guide rule-writing in or across different geographic districts, and new pseudo-legal debates emerging continuously. Though this may seem like a natural system to those already engaged with issues of content moderation from a legal perspective, the level of resources and investment required to participate in such a system might exclude less privileged groups of users from discussions of speech in ways that mirror the ways that public legal systems have consistently favored individuals and organizations with greater access to resources.

Structural and Functional Perspectives:

In his 2015 work, "The Virtues of Moderation", Grimmelmann [33] presented an initial taxonomy of the socio-technical approaches platforms take to moderation, including "organizing", "excluding", "pricing", and "norm-setting". Grimmelmann also provided a taxonomy of the different ways in which each of these can be performed, comparing, e.g., centralized versus decentralized moderation, automatic versus manual moderation, and ex post versus ex ante

moderation. Crawford and Gillespie focused on one specific moderation feature – the "flag" on social media, the tool through which users report content that they believe violates rules or norms [16]. 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. Blackwell et al. [6] detailed the consequences of this sort of classification – while it can validate users' experiences by making norms clear, it can also invalidate the experiences of users whose experiences do not match the classification scheme. Flagging mechanisms can also be gamed or abused; Crawford and Gillespie note cases where organized groups of users flagged content *en masse* as a form of attack on the content creators [16, p. 420-421]. Despite these vulnerabilities, the flag is core to moderation processes on most major social platforms; given the enormous volume of content that is produced, companies that use a centralized approach to moderation must rely partially on users' reports to identify which content to examine.

A related body of work has examined platforms' policies and the impact that they have on users. In attempting to better understand platforms' policies for dealing with harassment, Pater et al. examined various platform policy documents, from terms of service to community guidelines to parental and teen/youth guides [67]. 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 considered when determining whether actions would be defined as harassment.[5] In complementary work, West 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 [84]. Suzor et al. proposed specific ways in which increased transparency could help educate users and establish a sense of trust in these processes [81]. Achieving transparency is a significant challenge; increased transparency would allow for more public and perhaps democratic

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

debate about companies' processes, but could also highlight the many ways in which these companies are currently ill-equipped to make context-sensitive decisions.

A final body of work in the platforms and policies 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. However, extensive research by Sarah Roberts has uncovered the central role of human labor in what she terms "commercial content moderation" [71]. 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 primarily by algorithms and company employees.[6] Deep investigation by Roberts and work from several 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 [72, p. 50].

Sociotechnical Interventions on Centrally-Governed Platforms:

Though the above research points out real and serious issues in platform-based moderation, relatively little work has quantitatively analyzed the impact of platform moderation decisions at scale. Chandrasekharan et al.'s work studying the impact of Reddit's decision to ban certain forums is noteworthy in this regard. They found that the ban led to a decrease in behaviors previously characteristic of the excluded communities [11], though they did not analyze changes in problematic behaviors that were not directly associated with these communities. Acquiring sufficient data to fully evaluate platforms' decisions can be difficult, but work evaluating specific outcomes is important and necessary.

[6] Gillespie notes that, in its early days, Facebook relied on Harvard students to volunteer their time as moderators [32, p. 118].

Other research exploring potential improvements to moderation on centrally-governed platforms leverages users' collective effort. Geiger studied one collective approach to mutual moderation via Twitter blocklists, where users work together to coordinate lists of users who they all agree to block [26], a phenomenon further explored in research by Jhaver et al. [41], though in a sense this is more a form of community self-moderation than platform-driven moderation. Other work has proposed methods for making user reports more effective; Ghosh, Kale, and McAfee [30] proposed a computational approach for identifying trustworthy volunteer content raters and screening out bad actors, an approach that translates well to, e.g., identifying trustworthy reports on Reddit.

Beyond the above approaches, a popular field of recent study has been the development of algorithms to automatically identify and remove offensive content on platforms at scale. Work in this domain has focused largely on spaces like Twitter [7], Instagram [53], and online news comments [61].[7] These approaches face a number of significant challenges. First, there is no standard way to define problematic content, so these papers typically present a classification schema, a method for detection, and measures for evaluating these methods' effectiveness simultaneously. Each paper defines its focus in a slightly different way; Nobata et al. [61] focus on hate speech, but do not provide a specific definition by which raters applied this label. Burnap and Williams [7, p. 227] also focus on hate speech, defining it to raters as content that is "offensive or antagonistic in terms of race ethnicity or religion". Liu et al. [53, p. 183] detect "hostile" content, defined as "containing harassing, threatening, or offensive language directed toward a specific individual or group". These differing definitions make comparing the effectiveness of different approaches virtually impossible. The "problem of definition" is a major challenge in research on algorithmic content moderation.

Gröndahl et al. showed that automated detection systems are also very vulnerable to slight changes in text. In a 2018 paper, they tested seven "state-of-the-art" hate speech detection models, finding that each was only successful when tested on the same type of

[7]There is a somewhat different set of challenges in automated detection of problematic content in community self-moderated spaces, which I discuss in the following section.

data they were trained on [34]. They also showed that these algorithms were vulnerable to simple workarounds, such as inserting typos, shifting word boundaries, or adding innocuous words, and that even Google's "Perspective" API is vulnerable to such attacks. Binns et al. found that content moderation algorithms' decisions were significantly impacted by who labeled the training data, showing that these algorithms made different predictions when trained by, e.g., a subset of women raters compared to men raters [5]. Thus, these algorithms have the propensity to inherit the biases of their creators and can amplify them across a potentially-massive scale.

There are reasonable questions to be asked about the long-term viability of the models for moderation that centrally-governed platforms have adopted, as described in the work cited above. For example, the structure and (in)visibility of the labor of the contractors Roberts describes [72] suggests that platforms may see these contractors partly as a stopgap measure, which would be scaled back when algorithms reach a certain threshold of quality. While moderation algorithms have become quite effective in identifying spam, fake accounts, and explicit pornography,[8] their performance in identifying fake news or hate speech or cyberbullying is nowhere near as effective and, because of the complexity of these problems, may never be. It is thus reasonable to ask whether the future of speech on social platforms is inextricably tied to the labor of armies of commercial content moderators combined with (or replaced by) flawed algorithmic detection processes. In the following section, I consider another option.

2.2.2 The Communities Perspective: Users' Intra-group Moderation

In studying the processes of moderation in online communities, scholars including Viégas et al. [83], Forte, Larco, and Bruckman [23], and Kollock and Smith [48] have drawn on

[8] See, e.g., accuracy details provided in https://web.archive.org/web/20200204065414/https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works.

a framework proposed by Ostrom for "Design Principles of Long-Surviving, Self-Organized Resource Regimes" [64], [65, pp. 255–271]. Ostrom's eight principles were created based on her research in offline communities of up to 15,000 members that managed common pools of resources, per the classic 'Tragedy of the Commons' problem, including communities in Japan, Switzerland, Turkey, Sri Lanka, Nova Scotia, and the United States [63]. Ostrom's work explored governance systems that she explicitly framed as alternatives to firm and state-driven approaches to governance [63, pp. 8–20, 40–45], [66, pp. 10–13]. I argue, as did Ostrom, that there is "an alternative way" of addressing problems of community governance [63, p. 15] that has been largely overlooked in the public discourse. Whereas the platforms and policies perspective in moderation research works from the assumption that the professions and states will be at the core of moderation decisions and processes, a substantial body of research has studied platforms that allow users significant leeway to self-moderate. In this section I discuss this research in depth. Note that I do not argue here that self-moderation in online communities is a utopian ideal to be achieved. Self-moderation structures can be flawed in a wide variety of ways; the same structures that allow minority groups to create supportive spaces allow, e.g., white nationalist groups to create safe spaces of their own. The goal of this section is to explore some of the nuances of self-moderation in order to help push the discussion about the potential for these models in a more productive and generative direction.

Since the beginning of the internet, a large portion of social interactions online have been structured within online communities. Research studying community moderation began with a body of ethnographic work, beginning in the late 1970s (e.g., [38]) and peaking in volume in the mid-to-late 1990s. This work focused primarily on spaces like Usenet and MUDs[9] (e.g., [18, 37, 54, 69, 77, 79]), but also explored spaces like Electronic Bulletin Board Systems [70, pp. 131–144], [80], the Whole Earth 'Lectronic Link (WELL) [70, pp. 18–64], Lucasfilm's Habitat [58], and numerous other smaller spaces. These platforms were largely

[9](Multi-User Dungeons, Domains, or Dimensions)

decentralized, independently-operated, and relatively technically unsophisticated compared to modern platforms, but researchers identified a number of complex social processes for moderation, many of which incorporated concepts of "virtual democracy". Community moderation work in the 2000s and early 2010s focused more on "peer production" platforms [3, 2] including Wikipedia (e.g., [4, 23, 28, 46, 76]) and free and open-source software (FOSS) communities (e.g., [44, 62]). A smaller body of work in this era focused on "distributed" or "crowdsourced" moderation [51, 52], which would provide a foundation for later analysis of platforms like Reddit.[10] This era also saw an increase in research on "citizen governance" from a political science perspective, with analyses of moderation of public debate forums like Cornell's RegulationRoom [22, 60], the McGill Online Design Studio [22], and British government-run online discussion fora [87].

These platforms have fundamentally different characteristics than the major social media platforms that Gillespie [32], Klonick [47], and others have focused on. While platforms like Twitter, Instagram, YouTube, and Facebook (excepting Facebook Groups and Pages) are moderated almost exclusively by companies, the day-to-day moderation of community-centric platforms like Twitch, Wikipedia, and Reddit is handled primarily by users.

Frameworks for analysis of community moderation:

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. For example, in a wide-ranging synthesis of research on online communities published in 2012, Kraut and Resnick [50] 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. The importance of volunteer moderators' work has been clearly established. Many users prefer to participate

[10] See Kou et al.'s work [49] on "crowdsourced" moderation in League of Legends for another more recent example.

in spaces that are well-moderated [85], and thoughtful moderation can increase the quality of users' contributions [15] and help steer communities through periods of turbulence [45, 68]. Though platform administrators[11] do typically have "veto power" over volunteer moderators' decisions in that they can remove users, content, or entire communities without input from these volunteers, the relationship between platforms and volunteer moderators is typically distant or even nonexistent. Most volunteer moderators on these platforms never encounter interference from platform administrators in the way they moderate their communities [75, p. 11].

Though the practices of these moderators are diverse and complex, Seering et al. [75] propose a general framework that situates moderation processes within three different levels of granularity: (1) on an everyday, in-the-moment level, moderators interact with community members, warn potential offenders and explain rules, remove content and/or users when necessary, and deal with the fallout of these removals; (2) on a level that spans weeks or months, moderators learn how to moderate. This includes their processes of recruitment, role differentiation, learning how to handle various situations, and development of an overall moderation philosophy; and (3) on the broadest level, which spans the full lifetime of a community, moderators respond to internal community dynamics, platform developments, and cultural shifts by revising community rules and how they are enforced. Each of these processes is intricately intertwined with the others, with moderation incidents often impacting phases of all three process levels. I focus only on the first and third of these levels, as relatively little empirical research outside of Seering et al. [75] and Squirrel's work on Reddit moderation [78] has explored the second, but I discuss the second later in this chapter as an area that merits further research.

[11] I use the term "platform administrators" to refer to employees of companies like Facebook and Twitter who make and enforce final content moderation decisions.

The everyday labor of volunteer moderators:

In his 2016 work "The Civic Labor of Online Moderators" [57], Matias builds on literature from Gillespie [31], Shaw [76], Grimmelman [33], and others to enumerate a number of themes in volunteer moderation practice. First among these is the idea of "Moderation as Free Labor in the Social Factory of Internet Platforms" [57, p. 3], the idea that several major platforms rely extensively on free labor from users to operate. While this is a common argument in the context of content generation, it applies equally if not more so to moderation. As Gillespie argued in his *Custodians of the Internet*, "moderation is, in many ways, *the* commodity that platforms offer" [32, p. 13]. In the case of platforms like Reddit, it could be argued that *the* commodity that the platform offers is in fact not offered by the platform at all, but rather provided mostly by users to other users.

Each online community has its own goals that vary according to its age, size, needs, level of development, and other factors. Wohn identifies four (non-exclusive) roles moderators can play in Twitch communities: "Helping hands", "Justice Enforcers", "Surveillance Units", and "Conversationalists" [86, pp. 160: 6–7], and each of these terms implies a different set of goals. For example, Conversationalists aim to facilitate an active social environment with meaningful conversations, while Justice Enforcers aim to impose a specific set of norms or values, e.g., removing racist or sexist content. In work on political discussion groups, Epstein and Leshed identify three related but distinct goals for moderators in political discussion forums – keeping the discussion "in good order", collecting useful, quality content to be relayed to policymakers, and "building a community of civic-minded individuals" [20, pp. 4: 4–5]. Thus, the processes of community moderation and the approaches that moderators take vary widely and extend far beyond simple filtering and removal.

The mechanisms that moderators use for regulating behavior of new and established members have been studied in depth, with distinctions often made between social and sociotechnical approaches. The latter category is often more visible. For example, it is unlikely that a user could participate in public online groups for any significant length of time with-

out seeing at least an occasional ban or content removal. On some platforms, community members participate directly in the moderation process via "flagging" features, which can bring issues to moderators' attention. While Crawford and Gillespie focused on the use of flags to report problematic content to platforms [16], flagging and reporting tools on sites like Reddit are also widely used to send reports to communities' moderators, and moderators in busy spaces often rely significantly on user reports to focus their attention. However, these flagging tools can be abused just as easily in community-based social settings as on large-scale platforms. Many moderators, especially on Reddit, have to deal with waves of "report spam" where users report particular content *en masse* as an attempt to silence its creator or where users report a large number of reasonable posts in order to disrupt moderators' workflows by requiring them to dig through false reports. Because of issues like these and the broader challenges of managing large volumes of content, various work has identified the importance of usability in tools to manage this workflow [20, pp. 4: 10–12], [35].

Though a significant portion of moderators' work can require tools or features, Seering et al. [75] found that many moderators consider social approaches to be more important or central to their moderation practices than technical approaches. These social approaches span a variety of types of interpersonal engagement. For example, drawing from surveys of Twitch users, Cai and Wohn identified five approaches that moderators take to dealing with problematic behaviors: Educating, Sympathizing, Shaming, Humor, and Blocking [8, pp. 167–169]. Other work has identified similar strategies across a wide variety of platforms, including both social spaces [75] and spaces for political discourse [20, pp. 4: 14–17]. Moderators frequently use an escalating set of responses to problematic behaviors which begin as social responses and escalate into technical responses (i.e., time-outs or bans) [75]. In many cases, moderators first communicate to an offender that their conduct is inappropriate via a private message or brief response to their comment; on platforms like Reddit and Facebook Groups, these warnings and explanations often follow comment removal, and Jhaver, Bruckman, and Gilbert found that, on Reddit, explanations had a positive impact on future

participation of the offender in question [40]. If an offender continues to behave poorly, moderators either issue a stern, direct warning or a brief time-out. This is eventually followed by a ban from the space. Depending on the severity of the infraction or the nature of the offender, moderators sometimes skip steps in this process. For example, if an offender is perceived to be a bot rather than a human user, moderators may skip directly to banning it. Similarly, users who display extreme behaviors (e.g., aggressive racial slurs, rape threats) are often immediately banned [75, p. 17]. These social strategies are not new, and are implicit in much prior work on moderation (e.g., [77, 79]), but are also closely related to typical human approaches to overseeing groups or communities in any context.

Beyond responding to problematic behaviors, one of the most common everyday tasks in the work of volunteer moderators is handling newcomers, which has been a challenge since the early social web, per the CommuniTree example that opens this dissertation. In her studies of "MicroMUSE", performed in the early-to-mid 1990s, Smith [77, p. 148] identified a nuanced and evolving process for integrating newcomers into the community. Though initially more lax and open, this process changed several years into MicroMUSE's operation in response to conflicts that followed rapid growth in the community's membership. Server administrators restricted the commands that visitors could use to interact with other users and required that all new members receive "sponsorship" from two existing members after a period of socialization. A program for "mentorship" of these newcomers was also created, providing liaisons between newcomers and the main user-base. All of these processes appear in similar forms in modern platforms, though rarely in combination. 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.[12] "Followers-only mode" on Twitch requires users to have been present for a certain amount of time before they are allowed to post [75], and Automoderator settings on Reddit can also prohibit new users from posting in certain communities [39]. Wikipedia maintains an

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

"Adopt-a-user" mentoring program that pairs new users with more experienced Wikipedians [59].

Guiding and shaping a community:

The second theme that Matias notes is "Moderation as Civic Participation" [57, p. 3], which describes the types of leadership, governance, and management that occur in groups both online and offline. Self-moderation online typically operates on a volunteer basis, with moderators often taking significant time out of their lives to work what some describe as a "second job" [75, p. 12]. Moderators' online labor is comparable to work that has historically been done in a wide variety of offline environments; when people volunteer to host a gardening club, lead a homeowner's association, or organize a local chess tournament, they take on responsibilities and gain the power to make certain decisions. Volunteers in both offline and online communities often see the fun of participation or the social recognition they receive as their reward; these communities are meaningful to them, and in contributing their labor they also contribute to a broader social sphere [75]. Wohn focuses in depth on these types of civic motivations [86, p. 160:7–9], further underscoring the point that moderators contribute because spaces are meaningful to them, explaining how a "lack of appreciation" can be an emotional toll on moderators who want their community to appreciate the time and effort they put in.

The longer-term roles of moderators in shaping communities often include developing new rules and processes for moderation in a cyclical, evolving process. Sternberg, drawing from literature both in CSCW and the sociology of deviance, identified three rule-related social processes in online communities: *rule-breaking*, *rule-making*, and *rule-enforcement* [79, p. 155-169]. She noted that these processes take place in variable orderings. One might assume that rules are created for a space, and when users break or threaten to break these rules, moderators enforce them. However, as Sternberg notes, 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. Similarly, Seering et al. found that rule changes were often catalyzed by changes in internal community dynamics, often shaped by moderators' pre-existing values and occasionally by influence or intervention from platforms, and these new rules led to additional changes in internal dynamics [75, p. 16–19]. Building from work on Reddit, Squirrel [78] terms this back-and-forth between moderators and users a 'platform dialectic', where moderators deploy platform affordances to nudge users, users respond in sometimes unexpected ways, and moderators then re-deploy affordances.

As is the case with major social media platforms' moderation practices, volunteer moderators' decisions are frequently opaque and rarely involve community input [75]. No major modern social platforms have been designed with tools for making democratic decisions in moderation; there are no technical features on Facebook Groups, Reddit, Twitch, or Discord that facilitate election of moderators or votes on rule changes or formal referendums.[13] Accordingly, the final major theme that Matias identifies is "Moderation as Oligarchy" [57, p. 4]. Though oligarchic, dictatorial, or feudal approaches to moderation have been the default at least since the rise of Reddit, early online communities experimented with various models for more democratic moderation. MacKinnon [54] described a widely-cited incident originally reported by Dibbel [17] of community "justice" in a fantasy-themed MUD called LambdaMOO, which eventually led to the implementation of basic democratic mechanisms for moderation. Though this LambdaMOO incident is the most famous and most cited [17, 54], it was not an isolated case. Smith describes another incident in MicroMUSE where a teenage user named "Swagger" built an "Orgasm Room" filled with sex objects where he brought female players [77, 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 now happen regularly across numerous

[13] As Forte, Larco, and Bruckmam note, Wikipedia's decision-making processes could in some senses be called Democratic [23], but as the recruitment of new editors has slowed, Wikipedia has become decidedly more oligarchic [36, 76].

platforms without triggering backlash, in this case the residents of MicroMUSE revolted in defense of Swagger's perceived right to an opportunity to defend his actions, and two staff helped Swagger re-create his character. Though Swagger was eventually still permanently banned from the community, this revolt led to the organization of a community-wide town hall to discuss decision-making processes. 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, MicroMUSE adopted a new governing charter that created a "Citizens Council", established procedures for handling incidents of misbehavior, and provided a limited right of appeal [77, p. 148-158].

Forms of democratic moderation have been experimented with even in companies with a stronger profit motive. Habitat, one of the "first attempts to create a very large-scale, commercial, many-user, graphic virtual environment" [58, p. 273], was designed and managed by Lucasfilm Games, a division of LucasArts Entertainment Company and launched in the late 1980s. These designers took an approach that focused on enabling users to build their own social structures and experiences. For example, following a suggestion from a community member, these designers created a "Sheriff" role, implemented a voting system, and worked with community volunteers to hold an election. The community organized a public debate, where three candidates each made statements and answered questions. A vote was held, and one of the three candidates was elected. This Sheriff was initially only a figurehead, as they were granted no formal powers; The designers were uncertain what powers to grant the Sheriff, so eventually they decided to hold another community vote on several referenda about how the "legal system" ought to be set up. Though they were unable to act on the results of the referenda, as the version of the system in which these events took place was shut down shortly after voting took place [58, pp. 290–291], their process showed potential for community participation in moderation even on commercial platforms.

The processes of rule-development in individual online communities frequently mirror the processes described by Gillespie [32] that occur as whole platforms develop, but typically

on a much smaller scale. The philosophies moderators draw on for rule-writing are often a combination of personal values and prior experience within online spaces, but these rules are frequently challenged by incidents that push moderators to re-evaluate their stances, just as Facebook, Twitter, and others have been forced to respond to a constant stream of unanticipated incidents. In practice, rule-writing often comes down to a mix between idealism and pragmatism. As Juneja, Ramasubramanian, and Mitra [42] note, moderators have mixed opinions about being transparent in both their rule-writing processes and enforcement processes. While some feel that transparency is important both as a general ethical principle and as a way of helping users understand what behavior is acceptable, others feel that transparency can lead to greater abuse. Research studying user behaviors supports both of these points of view; one body of work has found that providing explanations of removals is effective in reducing subsequent misbehavior [40], while other work supports this latter idea, at least in a networked social media context, showing that users are persistent in finding ways around word filters when they can figure out how those filters work [9, 29].

As noted above, though volunteer community moderators' processes often mirror platforms' processes, moderators on community-based platforms rarely experience direct interference from platform administrators into how they run their communities [75]. This could be interpreted to mean that these platforms recognize the right of the community moderators to create and enforce their own rules, but the reality is less straightforward. No major modern social platform companies explicitly cede any final moderation authority to volunteer community moderators on their platforms. Reddit, for example, has as a company traditionally been very reluctant to take large-scale action [56, p. 340], but in extreme cases it has banned entire communities from the site without potential for appeal. Facebook has also banned groups in varying circumstances, and can take action against individual posts within communities as well. Though these companies permit volunteer moderators to handle the vast majority of moderation decisions, they typically reserve the right to make a final decision in the rare cases where moderators are not managing communities to the platforms'

satisfaction.

Volunteer community-based moderation and moderation performed by platform administrators could be seen as two different layers of an organization, and they each perform a set of relatively distinct tasks; while volunteer moderators on Reddit handle most of the day-to-day moderation, Reddit as a platform uses internal metrics, data, and statistics to track covert political influence campaigns and bans offending accounts. However, these two "layers" are very disconnected. Seering et al. found that, while these two types of moderators were often working toward the same goals, communication between them is rare [75]. Platform administrators tend not to share their internal data or their goals with volunteer user moderators, and the two groups almost never collaborate directly on a specific problem in any structured way.

Social and Technical Interventions in Community Moderation

In discussing algorithmic approaches to platform-level moderation above, I noted the "problem of definition", where the (understandable) inability of researchers to coalesce on a single definition for problematic behaviors makes it difficult to compare the effectiveness of different algorithmic approaches. Algorithmic approaches have also been proposed for community self-moderated spaces, but the flexibility of these spaces allows researchers to take new approaches to these problems. Chandrasekharan et al. avoid defining what behaviors are "problematic" by using external communities' definitions of problematic behavior to define rules for a new community. This results in a classifier that works not from consciously defined rules but from the aggregation of prior users' moderation decisions [10]. Chandrasekharan et al. take an important step by choosing to use this classifier in a way that brings comments to the attention of moderators rather than removing the comments without human oversight. However, even this type of classifier is vulnerable to bias; if not used carefully, it can implicitly impose majority norms on minority communities. It is also important to be wary of taking any decision-making away from moderators; the human decision-making processes

that drive shifts in rules, frequently via conversations between moderators, are core to the evolution of communities [75, pp. 16–21]. Automated moderation is a delicate and perhaps even dangerous approach because of the way it can subtly shape moderation at scale in ways that supplant human consideration. However, it remains a necessary approach, particularly in large communities, and careful, context-aware research in this domain is important. As argued in work from Jhaver et al. [39] and Seering et al., it is important to develop tools that "support, rather than supplant, the judgment of users" [75, p. 3], and, per Geiger [27], when analyzing algorithms of this level of complexity researchers must not simply "open up the black box", but rather should go further to examine algorithms' impact on sociotechnical processes.

Though various platform-level interventions have been proposed for improving moderation processes, e.g., algorithmic detection tools, the flexibility and diversity of online communities has led to a wider variety of potential sociotechnical interventions in these spaces. These interventions take various forms. For example, one type of intervention leverages social support and the help of friends in responding to harassment or other problematic behaviors. Mahar, Zhang, and Karger [55] 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. Blackwell et al. [6] 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. These practices are already in place informally in many contexts; visible figures in Twitch communities often designate moderators to pre-screen chat messages in a way similar to the email screening in *Squadbox*, and moderators frequently find it useful across multiple types of communities when users band together to defend the community from malicious users.

Though most research on interventions has focused on *reactive* approaches to moderation – those that remove or filter problematic behaviors after they have occurred – another small body of research has begun to explore more *proactive* approaches to encourage people

to behave well in the first place. Preliminary work has explored using interface elements (e.g., CAPTCHAS) designed based on principles from psychology, finding some success in encouraging more positive and thoughtful comments [73]. Other recent work has employed "empathy nudges" to attempt to encourage bystander intervention, with mixed results [82]. A particularly new line of work has even looked at the use of community-embedded chatbots to help strengthen community identity and clarify norms, but this work has not yet been empirically tested beyond initial exploratory work [74]. Among the potential research directions for interventions, this space is perhaps the most wide-open; the design space for tools that proactively shift community behaviors in more positive directions is constrained mostly only by the creativity of interested researchers.

Bibliography

[1] Jan Philipp Albrecht. How the gdpr will change the world. *European Data Protection Law Review*, 2(3), 2016.

[2] Yochai Benkler. Peer production and cooperation. In Johannes M Bauer and Michael Latzer, editors, *Handbook on the Economics of the Internet*, pages 91–119. Edward Elgar Publishing, Cheltenham, United Kingdom, 2016.

[3] Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. Peer production: A form of collective intelligence. In Thomas Malone and Michael Bernstein, editors, *Handbook of Collective Intelligence*, pages 175–204. MIT Press, Cambridge, MA, USA, 2015.

[4] Matt Billings and Leon A. Watts. Understanding dispute resolution online: Using text to reflect personal and substantive issues in conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1447–1456, New York, NY, USA, 2010. ACM.

[5] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[6] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):24:1–24:19, December 2017.

[7] 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 & Internet*, 7(2):223–242, 2015.

[8] Jie Cai and Donghee Yvette Wohn. What are effective strategies of handling harassment on twitch? users' perspectives. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, pages 166–170, New York, NY, USA, 2019. ACM.

[9] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1201–1213, New York, NY, USA, 2016. ACM.

[10] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[11] 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. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):31:1–31:22, December 2017.

[12] Danielle Keats Citron. *Hate Crimes in Cyberspace*. Harvard University Press, Cambridge, MA, USA, 2014.

[13] Danielle Keats Citron and Benjamin Wittes. The internet will not break: Denying bad samaritans sec. 230 immunity. *Fordham L. Rev.*, 86:401, 2017.

[14] Danielle Keats Citron and Benjamin Wittes. The problem isn't just backpage: Revising section 230 immunity. *Georgetown Law Technology Review (2018)*, 2:453–473, July 2018.

[15] Dan Cosley, Dan Frankowski, Sara Kiesler, Loren Terveen, and John Riedl. How oversight improves member-maintained communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 11–20, New York, NY, USA, 2005. ACM.

[16] 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.

[17] 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*, December 23:36–42, 1993.

[18] 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, UK, 1st edition, 1999.

[19] Evelyn Douek. Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility. *N.C. J.L. & Tech*, 21:1–78, 2019.

[20] Dmitry Epstein and Gilly Leshed. The magic sauce: Practices of facilitation in online policy deliberation. *Journal of Public Deliberation*, 12(1), 2016.

[21] European Commission. Regulation (eu) 2016/679 (general data protection regulation). oj l 119, 04.05.2016; cor. oj l 127, 23.5.2018., 2016.

[22] Cynthia Farina, Hoi Kong, Cheryl Blake, and Mary Newhart. Democratic Deliberation in the Wild: The McGill Online Design Studio and the Regulation Room Project. *Fordham Urb. L.J.*, 41:1527, 2014.

[23] Andrea Forte, Vanesa Larco, and Amy Bruckman. Decentralization in Wikipedia Governance. *Journal of Management Information Systems*, 26(1):49–72, 2009.

[24] Mary Anne Franks. "Revenge Porn" Reform: A View from the Front Lines. *Fla. L. Rev.*, 69:1251–1337, 2017.

[25] Mary Anne Franks. *The Cult of the Constitution*. Stanford University Press, Palo Alto, CA, USA, 2019.

[26] R. Stuart Geiger. Bot-based collective blocklists in twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803, 2016.

[27] R Stuart Geiger. Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society*, 4(2), 2017.

[28] R. Stuart Geiger and David 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*, CSCW '10, pages 117–126, New York, NY, USA, 2010. ACM.

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

[30] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 167–176, New York, NY, USA, 2011. ACM.

[31] Tarleton Gillespie. The politics of 'platforms'. *New Media & Society*, 12(3):347–364, 2010.

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

[33] James Grimmelmann. The Virtues of Moderation. *Yale J.L. & Tech*, 17:42–109, 2015.

[34] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, pages 2–12, New York, NY, USA, 2018. ACM.

[35] David Gurzick, Kevin F. White, Wayne G. Lutters, and Lee Boot. A view from mount olympus: The impact of activity tracking tools on the character and practice of moderation. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, GROUP '09, pages 361–370, New York, NY, USA, 2009. ACM.

[36] 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.

[37] 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.

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

[39] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5):31:1–31:35, July 2019.

[40] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[41] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.*, 25(2):12:1–12:33, March 2018.

[42] Prerna Juneja, Deepika Ramasubramanian, and Tanushree Mitra. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. In *Proceedings of the 21st International Conference on Supporting Group Work*, New York, NY, USA, 2020. ACM.

[43] David Kaye. *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports, New York, NY, USA, 2019.

[44] Christopher M Kelty. *Two bits: The cultural significance of free software*. Duke University Press, Durham, NC, USA, 2008.

[45] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. Surviving an "eternal september": How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1152–1156, New York, NY, USA, 2016. ACM.

[46] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 453–462, New York, NY, USA, 2007. ACM.

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

[48] Peter Kollock and Marc Smith. Managing the virtual commons: Cooperation and conflict in computer communities. In Susan Herring, editor, *Computer-mediated Communication: Linguistic, Social, and Cross-cultural Perspectives*, pages 109–128. John Benjamins Publishing, Amsterdam, Netherlands, 1996.

[49] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. Managing disruptive behavior through non-hierarchical governance: Crowdsourcing in league of legends and weibo. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):62:1–62:17, December 2017.

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

[51] Cliff Lampe and Paul Resnick. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 543–550, New York, NY, USA, 2004. ACM.

[52] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317–326, 2014.

[53] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, volume 91, pages 181–190, Menlo Park, CA, USA, 2018. AAAI.

[54] Richard MacKinnon. Virtual rape. *Journal of Computer-Mediated Communication*, 2(4):1–2, 1997.

[55] Kaitlin Mahar, Amy X. Zhang, and David Karger. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 586:1–586:13, New York, NY, USA, 2018. ACM.

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

[57] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2), 2019.

[58] Chip Morningstar and F Randall Farmer. The Lessons of Lucasfilm's Habitat. In Michael Benedikt, editor, *Cyberspace: First Steps*, pages 273–301. MIT Press, Cambridge, MA, USA, 1991.

[59] 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*, WikiSym '11, pages 173–182, New York, NY, USA, 2011. ACM.

[60] Chaebong Nam. Behind the interface: Human moderation for deliberative engagement in an eRulemaking discussion. *Government Information Quarterly*, 2019.

[61] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[62] Siobhán O'Mahony and Fabrizio Ferraro. The emergence of governance in an open source community. *Academy of Management Journal*, 50(5):1079–1106, 2007.

[63] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK, 1990.

[64] Elinor Ostrom. Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3):137–158, September 2000.

[65] Elinor Ostrom. *Understanding institutional diversity*. Princeton university press, Princeton, NJ, USA, 2005.

[66] Elinor Ostrom. *The Future of the Commons*. Institute of Economic Affairs, London, England, UK, 2010.

[67] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, GROUP '16, pages 369–374, New York, NY, USA, 2016. ACM.

[68] David J Phillips. Defending the Boundaries: Identifying and Countering Threats in a Usenet Newsgroup. *The Information Society*, 12(1):39–62, 1996.

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

[70] Howard Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier*. Addison Wesley Publishing Company, Boston, MA, USA, 1993.

[71] Sarah T. Roberts. Commercial Content Moderation: Digital Laborers' Dirty Work. In Safiya Umoja Noble and Brendesha M. Tynes, editors, *The Intersectional Internet: Race, Sex, Class and Culture Online*, pages 147–160. Peter Lang Digital Formations series, New York, NY, USA, 2016.

[72] Sarah T Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, CT, USA, 2019.

[73] 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 606:1–606:14, New York, NY, USA, 2019. ACM.

[74] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. It takes a village: Integrating an adaptive chatbot into an online gaming community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, 2020. ACM.

[75] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[76] 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.

[77] 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, New York, NY, USA, 1st edition, 1999.

[78] Tim Squirrell. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society*, 21(9):1910–1927, 2019.

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

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

[81] 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*, 13:1526–1543, 2019.

[82] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[83] Fernanda B Viégas, Martin Wattenberg, and Matthew M McKeon. The Hidden Order of Wikipedia. In Douglas Schuler, editor, *Online Communities and Social Computing*, pages 445–454, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

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

[85] 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, 10 2006.

[86] Donghee Yvette Wohn. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 160:1–160:13, New York, NY, USA, 2019. ACM.

[87] Scott Wright. Government-run Online Discussion Fora: Moderation, Censorship and the Shadow of Control1. *The British Journal of Politics and International Relations*, 8(4):550–568, 2006.

3

Organizational Perspectives: Cooperative Responsibility and Content Moderation

3.1 Introduction

The previous chapter explored the various research perspectives that have contributed to the understanding of content moderation. In this chapter, I explore a question that emerges from the intersection of those perspectives – "How can the relationship between volunteer community moderators and platform administrators be made more productive?" – in order to give a high-level perspective on moderation processes before I focus on specific social interactions in subsequent chapters. I take a deep dive into the concept of "cooperative responsibility", which was briefly mentioned in the previous chapter's discussion of this question.

As discussed in the previous chapter, the processes for moderating user-generated content online are complex and vary across different types of companies. On the platforms that will be discussed at length in this dissertation, volunteer community moderators screen content posted by other users, and, when a piece of content is determined to be inappropriate, these moderators can take a variety of actions ranging from social interventions (i.e., warning a

user) to removing them and/or the content from the community. While these platforms also screen individual pieces of content, they also focus more broadly on questions of whether specific communities should be permitted to remain on the site. While the work cited in the previous chapter has been valuable in providing foundational knowledge of processes either from the perspective of platforms or volunteer community moderators, a full understanding of moderation on platforms like Reddit, Facebook Groups, and Twitch cannot be achieved without analyzing the integration of these pieces. Moderation on, e.g., Reddit cannot be fully understood through separate analyses of volunteer moderators' processes and Reddit platform processes; analyses of the ways these processes interact are necessary. The approaches that platforms take to moderation can impact users' experiences [22] and can shape how volunteer moderators' approach their work [19], and the ways in which users interact with moderation policies can shape platforms' approaches in return [6]. In an ecosystem where "moderation is, in many ways, *the* commodity that platforms offer" [8, p. 13], each piece of the process that creates this commodity must be considered in relation to the others.

The primary question I attempt to answer in this chapter is, "to what extent have the systems of content moderation of Reddit, Twitch, and Facebook Groups achieved each of the four steps proposed toward the standard of *cooperative responsibility* in governance?" In their 2018 work, "Governing online platforms: From contested to cooperative responsibility", Helberger, Pierson, and Poell discuss how responsibility is allocated between platforms and users across a variety of online platforms from Uber to Facebook to Coursera and more, particularly in the domain of governance [10]. They argue that, per the concept they name "cooperative responsibility", "platforms and users need to agree on the appropriate division of labor with regard to managing responsibility for their role in public space" [10, p. 2]. Helberger, Pierson, and Poell recognize that this is not a simple task, naming several significant obstacles to cooperation in agreeing on appropriate division of labor. First, platforms are typically transnational corporations that must make decisions with global impact, but their data is typically stored and processed in the US and as such they must attend to US

circumstances. Second, platforms are structured in complex and frequently opaque ways, with "black-box" algorithms that can conceal decision-making processes. Finally, there are complex power dynamics between users and the platform, and both the platform and its users already play roles in shaping values and norms in the space. Though operationalizing cooperative responsibility is challenging and perhaps impossible in its fullest sense, the concept remains useful as a lens though which moderation structures can be analyzed.

The ecosystems that Helberger, Pierson, and Poell explore when discussing the applications of cooperative responsibility do not include a case where distinctions are drawn between categories of users that mirror the distinctions between volunteer moderators and regular users; while in one of their examples they separate out Uber drivers from Uber riders, and both are "users" of the Uber app, Uber drivers have a clear, formalized relationship with the company. However, even though this type of scenario is not directly discussed by Helberger, Pierson, and Poell, the core framing of cooperative responsibility still applies well. Per their description, the concept of cooperative responsibility is derived in some part from the "problem of many hands" [20], a situation where many stakeholders are involved both in creating and resolving a problem "in a manner that makes it difficult to identify who is responsible for which actions and what consequences and on this basis [to] allocate accountability and responsibility accordingly" [10, p. 3]. This description fits the content moderation ecosystem, where there are many stakeholders, each with different capabilities, and a complex set of problems. Volunteer moderators and regular users each contribute in different (though sometimes overlapping) ways to these problems, and each has different tools to contribute to resolving them.

3.2 Cooperative Responsibility in Online Governance

Though Helberger, Pierson, and Poell do not explicitly state criteria to determine whether platforms and users have reached a state of cooperative responsibility, they do identify four

possible steps in the process to "organize the (re)distribution of responsibilities":

1. "to collectively define the essential public values at play in particular economic activities and modes of public exchange"

2. "for each stakeholder (platforms, governments, users, advertisers, and others) to accept that they have a role to play in the realization of these values"

3. "to develop a (multi-stakeholder) process of public deliberation and exchange, in which agreement can be reached between platforms, users, and public institutions on how important public values can be advanced"

4. "to translate the outcome of public deliberation and agreements into regulations, codes of conduct, terms of use and, last but not least, technologies (e.g. 'by design')" [10, p. 10].

The core of Helberger, Pierson, and Poell's paper focuses on demonstrating the need for a redistribution of responsibility by presenting three scenarios where redistribution might facilitate a better approach to specific problems. As the goal of this chapter is to present evidence that shows to what extent each of these steps has begun or has been completed, these steps must be defined in sufficient detail to facilitate evaluation in this context. However, while the above four steps are presented as a general framework, they are not expanded upon in depth outside of implicit connections to the three scenarios. Thus, in order to use them as an analytical tool, I propose an additional level of detail for each step that clarifies how each step might work in the context of content moderation on spaces that rely in part on community self-governance, and in particular as they apply to relationships between platform administrators, volunteer moderators, and regular users: The first step, collective definition of values, can be thought of as arrival at a mutual understanding about what a platform is *for*, which includes both the core values that are meant to govern interactions on the platform and the purpose for the creation of the platform, i.e., what types of content

it is meant to host. Values can be directly codified or inferred from rules or collectively understood and verbalizable norms. Collective definition of values has been achieved if users and moderators understand the platform administrators' values and vice versa. The second step is closely related – moderators and platform administrators must come to a collective agreement on what each of their roles is in participating in the space. The third step is arrival at mechanisms and processes for public deliberation and discussion about issues in content moderation. This step is, in part, the formalization of the ongoing process for maintaining steps one and two. The final step is to take the results of this deliberation and translate them into both things like policies and community guidelines, but also into the design of platform features that facilitate different approaches to moderation.

3.3 Methods

In order to better understand the relationship between volunteer moderators and platforms, I performed semi-structured interviews of 56 volunteer community moderators. I interviewed each of these moderators a first time between Fall 2016 and Spring 2018, and conducted follow-up interviews From Fall 2018 through Spring 2019 with 23 of them.[1] I interviewed 21 moderators from Reddit (9 follow-ups), 20 from Twitch (7 follow-ups), and 15 from Facebook Groups (7 follow-up interviews).[2] Interviews ran approximately 30 minutes on average, and interviewees were compensated 15 USD (or foreign equivalent) for participating. I initially recruited moderators through a combination of direct messaging moderators and snowball sampling, and follow-up interviewees were recruited by directly messaging past interviewees.

The first round of interviews focused on volunteer community moderators' processes in managing their communities, including social engagements with community members, use of

[1] I attempted to re-contact all first-round interviewees, and 23 out of 56 (approximately 40%) eventually participated in follow-up interviews. I feel that this is a strong response rate given the long period of time between first contact and follow-up.

[2] I have previously published an analysis of the results from the 56 first-round interviews [19], but have not published analysis that included the follow-up interviews. The original publication focused mostly on moderators' intra-community practices, while this publication analyzes the full system of moderation.

tools and automated detection systems, and development of rules and norms over time.While my results painted a strong picture of labor within communities, I did not have enough data to draw conclusions about organizational relationships between community moderators and platform administrators. In my follow-up interviews I focused in more depth on these relationships, asking moderators about their interactions with admins (if any), how they felt the platforms' rules impacted their communities, how the new tools and interface changes made by the platforms in the years between interviews impacted their moderation processes.

Each of the three platforms has different features, different rules and norms, and different official (and unofficial) moderation processes, and I note these differences where appropriate, but my goal is to present broader insights about organizational structures of platforms that rely in part on community self-moderation. Despite this broad focus, my analysis offers a limited perspective in part because I did not interview platform administrators. Given that I am, in part, analyzing the ways in which users and platforms negotiate division of labor, the actions taken by platform administrators are important to document. However, the type of agreement that Helberger, Pierson, and Poell describe in their concept of cooperative responsibility requires a basic level of communication and mutual understanding between parties [10], so community moderators' accounts of their interactions with platform admins should identify any forms of cooperative responsibility that do exist. I note, however, that this data cannot tell us about the actions taken by platform admins that are not publicly visible, and, if cooperative responsibility does not exist, can offer only a biased explanation of *why*. In order to supplement my interview data and provide at least a preliminary window into the values of the three companies, I reference official announcements, statements, and documents published by the companies that relate to moderation in these spaces. These documents include Reddit posts made by Reddit administrators, primarily Reddit CEO Steve Huffman, and public replies to users' comments by these administrators; versions of the Twitch community guidelines and official company blog posts about moderation on Twitch; and posts and announcements by Mark Zuckerberg and other high-ranking Facebook

employees that discuss moderation in ways that relate to Groups. I do not claim that this is a systematic review of all public statements made by these companies, but rather that referencing some of these documents can complement analysis of interviews with moderators.

My analysis is influenced significantly by Forte, Larco, and Bruckman [3] and Viégas, Wattenberg, and McKeon [21], both of which used qualitative data to analyze how a platform fits an existing organizational framework. In both of these works, they focused on Ostrom's design principles for self-organizing communities that manage natural resources [15, 16]. I aim to produce a "thick description" [5] of how moderation on Facebook Groups, Reddit, and Twitch fit within the frameworks proposed by Helberger, Pierson, and Poell [10]. Per Geertz, a "thick description" is one that allows someone outside the context being described to derive meaning from the evidence being presented, as opposed to a "thin description", which simply provides a series of facts. As such, my analysis processes followed the principles proposed in the cooperative responsibility framework, attempting to find evidence of whether and how moderation on these platforms fits within the steps described above. In order to do this, I analyzed each of the 56 initial interviews and 23 follow-up interviews and separated out chunks of each transcript that related to each of the four steps from Helberger, Pierson, and Poell. I identified a total of 91 relevant chunks. Table 3.1 shows the distribution of these chunks across the different platforms and steps.

Table 3.1: Number of quotes by platform and step in the cooperative responsibility process

	Facebook	**Reddit**	**Twitch**
1. Values	4	16	9
2. Roles	5	15	10
3. Process	16	11	2
4. Implementation	2	1	0

Note that I did not elect to calculate inter-rater reliability in my qualitative coding process. Per the explanation in McDonald, Schoenebeck, and Forte [14, pp. 72:13–14], the goal of this process was to gather information as it fit into an existing framework, not to evaluate or create a new framework. My process in this regard was again similar to the

processes in Viégas, Wattenberg, and McKeon [21] as well as Forte, Larco, and Bruckman [3].

3.4 Analysis

In order to understand how well the framework of cooperative responsibility applies to organizations of moderation, I present evidence that shows to what extent each of these steps has begun or has been completed. I proceed by examining each of the three platforms from which I interviewed moderators – Reddit, Twitch, and Facebook – and assess each according to my interpretation of the above four criteria.

3.4.1 Reddit and Cooperative Responsibility

Reddit has long prided itself as a space for free expression and open discourse, with an endless variety of communities centered around niche interests. Though Reddit originally emerged from "geek culture" [13], with popular spaces focused on science, technology, and gaming, Reddit has over time attracted a much broader audience. It has also grown in prominence in the space of political discourse, leading to long-lasting debates over whether certain types of extreme communities should be permitted on the site (e.g., "incel", or "involuntarily celibate" communities, white supremacist communities, extreme alt-right political communities). As the platform has drawn in a broader user base and become host to a wider variety of content, it has struggled to figure out how its identity will evolve in turn. However, it has been more open than the other platforms in discussing these questions directly with users.

Collective definition of essential public values:

Reddit was founded on cyberlibertarian values and as a space for people with deep connections to a somewhat niche set of interests. Reddit is, by its nature, a space for discussion; its technical structure is explicitly conversational, with text as the primary medium for com-

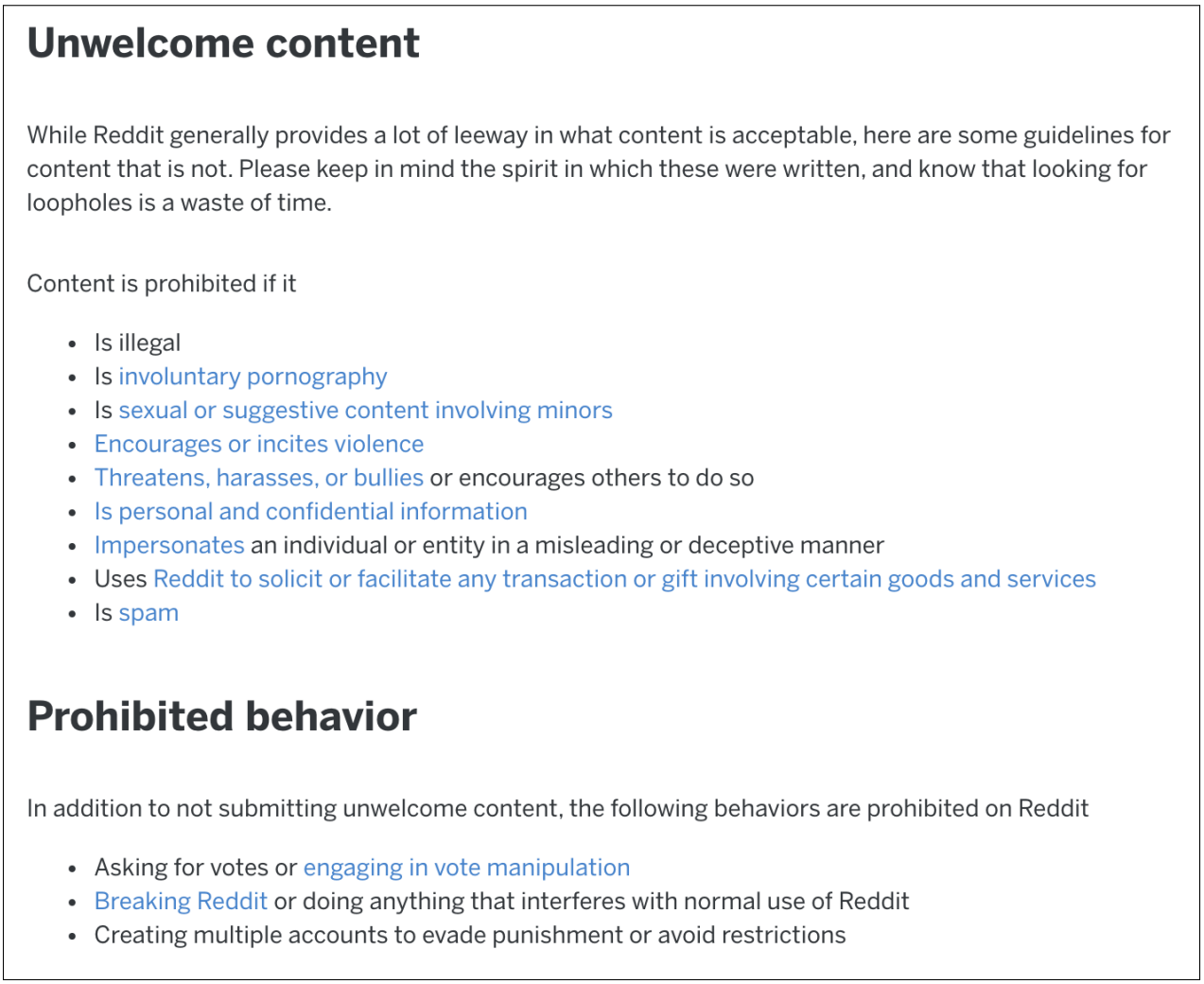
munication and the vast majority of communication on the site existing in threaded reply trees. The ethos and structure of Reddit thus made it a natural home for the sorts of geek cultures that prided themselves in "rational" discourse [13]. Thus, a core principle of Reddit ethos has long been the idea that 'good speech trumps bad speech' and that, given a platform for open discussion, better, more rational ideas will eventually triumph over worse ideas. This philosophy has slowed the reactions of Reddit administrators to problematic communities. Though Reddit was founded in 2005, the first prominent ban of a subreddit – /r/jailbait – was in 2011. This subreddit was dedicated to sexualized images of underage girls, and was only banned after an exposé by CNN's Anderson Cooper.[3] Other subsequent bans were made very reluctantly, including the ban of /r/n*ggers in 2013, which was not banned for racist content but rather the fact that its members had spread this behavior to other subreddits with the intent to disrupt them.[4] Several decisions to close major hate subreddits were justified similarly, with the ban of /r/BeatingWomen in 2014 made because users were sharing other users' personal information and collaborating to evade bans, not because of their posting and glorifying graphic imagery of women being beaten.[5] The subreddits /r/fatpeoplehate and /r/CoonTown (a successor to /r/n*ggers) were banned in 2015 [1], and recent years have seen an increase in subreddit bans focusing on subreddits that host extreme misogyny, repeated unmoderated calls for and glorification of violence, and illegal content. Subreddits like /r/Physical_Removal (which advocated for the deportation and or killing of liberals), /r/Incels (which frequently hosted content that glorified violence against women), and /r/WatchPeopleDie (which was dedicated to videos of real-life deaths of people notably including videos of the New Zealand mosque shooting in 2019) were all banned in the period of mid-2017 through mid-2019 after an expansion and clarification of the Reddit policy against glorifying violence.

[3] https://web.archive.org/web/20200430143329/https://www.dailydot.com/society/reddit-r-jailbait-shutdown-controversy/

[4] https://web.archive.org/web/20170423154121/https://www.theatlantic.com/technology/archive/2013/07/does-anything-go-the-rise-and-fall-of-a-racist-corner-of-reddit/277585/

[5] https://web.archive.org/web/20140621004602/http://www.dailydot.com/news/reddit-beating-women-banned/

Though Reddit does not have a formalized process for collectively defining "essential public values", there are existing lists of values defined both by the platform and the users. The strongest statement of the values of Reddit as a company can be seen in the Content Policy, which includes a brief list of types of content and behavior that are not permitted on Reddit (see Figure 3-1). Most of the listed items link to another page which includes one or two paragraphs explaining the rule in more depth.



Unwelcome content

While Reddit generally provides a lot of leeway in what content is acceptable, here are some guidelines for content that is not. Please keep in mind the spirit in which these were written, and know that looking for loopholes is a waste of time.

Content is prohibited if it

- Is illegal
- Is involuntary pornography
- Is sexual or suggestive content involving minors
- Encourages or incites violence
- Threatens, harasses, or bullies or encourages others to do so
- Is personal and confidential information
- Impersonates an individual or entity in a misleading or deceptive manner
- Uses Reddit to solicit or facilitate any transaction or gift involving certain goods and services
- Is spam

Prohibited behavior

In addition to not submitting unwelcome content, the following behaviors are prohibited on Reddit

- Asking for votes or engaging in vote manipulation
- Breaking Reddit or doing anything that interferes with normal use of Reddit
- Creating multiple accounts to evade punishment or avoid restrictions

Figure 3-1: Unwelcome content and prohibited behavior from Reddit's content policy, June 27th 2020.

This list is notably short, when compared with the rules of other platforms. But many of these rules are well known across Reddit. Most interviewees were familiar with the rules against illegal content, sexual or suggestive content involving minors, personal and confi-

dential information (i.e. "doxxing"), spam, and vote manipulation. Most moderators also mentioned that they thought there was a rule against self-promotion, but this is not an explicit rule in the content policy. Moderator interviewees were mixed as to how much they felt these policies impacted their subreddits' rules. One moderator used the content policy as a starting point:

> *"I think our sub rules not only incorporate Reddit content policy but are also more strict. So if I'm enforcing the sub rules, I'm probably automatically incorporating the site rules as well."* – R7

Another said that the sitewide policy informs the basis for all of their rules (R11), and that their community had chosen not to add additional rules at all. Perhaps the most common response, though, was for moderators to say that they didn't pay too much attention to the content policy:

> *"The overall guidelines don't affect us too much. The nature of an academic subreddit is that it's gonna follow the rules already unless its used for malicious reasons."* – R18

These content policies can thus be seen as a minimum set of standards for participation on Reddit.

There also exists a list of values created by users, which is called *Reddiquette*[6]. Whereas the formal Reddit Content policy outlines what content and conduct is permissible or prohibited on the site, Reddiquette is a less formal set of guidelines put together by users for how to behave on the site. The Reddiquette document includes suggestions such as "Read the rules of a community before making a submission", "Actually read an article before you vote on it (as opposed to just basing your vote on the title)" and "[don't] Write titles in ALL CAPS". Though these are suggestions and are not formally enforced by Reddit, most moderator interviewees referenced values that they use in moderating that align well with

[6]http://web.archive.org/web/20200623165026/https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/reddiquette

Reddiquette, and several specifically mentioned it. One interviewee said that their moderation team allows users to evaluate when Reddiquette has been breached – *"We allow reddiquette to reinforce itself with the up/downvotes"* (R11) – meaning that the moderators do not explicitly consider Reddiquette a set of rules by which they moderate, but they expect community members to vote on posts and comments in part based on the norms laid out in Reddiquette.

Overall, interviewees had a solid sense of most of the content policies and agreed in the abstract on core points of Reddiquette, but nearly all of them mentioned times when public debate had sprung up in response to issues that were not clearly covered by these, or when they felt that Reddit administrators had not applied these rules fairly and consistently.

> *"I think the, the number one problem that Reddit has had with the quarantining and banning of, um, subs is that they have not been consistent with their application of their own rules. So I think the answer is you have a consistent set of rules for what is acceptable and what's not acceptable and you apply them consistently."*
> – R6 Reinterview

Reddit engages in more open and direct communication and Q&A with users than most platforms. Administrators (notably including Reddit CEO Steve Huffman) post explanations when large-scale moderation actions are taken, and frequently respond to users' critiques voiced in the comments on those posts (albeit not always in a way that fully satisfies these users). Annual transparency reports[7] are shared with the community, which provide statistics about what content has been removed in the past year and why. While these reports provide some post-hoc transparency, they do not provide an ongoing window into Reddit's deliberative processes. More frequent updates from administrators related to moderation are posted the subreddit /r/ModNews, and many Reddit employees are active on this subreddit to answer questions. This, however, is a set of practices for transparency rather than a col-

[7] e.g., https://old.reddit.com/r/announcements/comments/f8y9nx/spring_forward_into_reddits_2019_transparency/

lective process for value definition. Such a process would require more active participation and agency by users than the current system, which allows users to give feedback that may or may not be considered.

Stakeholders' acceptance that they have a role:

The roles of volunteer moderators vs Reddit administrators have long been implicitly understood, at least in a general sense.

> *"Individual communities on Reddit may have their own rules in addition to ours and their own moderators to enforce them. Reddit provides tools to aid moderators, but does not prescribe their usage."* – Reddit Content Policy

This statement, in combination with the other pieces of the Content Policy, suggests that moderators have the right to oversee and manage their communities autonomously, largely independent from administrator intervention, so long as their communities do not become seriously problematic as a whole. If their community strays beyond the behavior boundaries set by the platform, they will then receive attention from administrators and potentially reduced autonomy until the problems are resolved.

> *"I think that as a moderator, our role should be more than simple you know, green light, red light on posts. Uh, I think a moderator is a, is a steward of their community."* – R11 reinterview

Though this idea might suggest that moderators seek to maintain their independence, moderators mentioned numerous cases where situations arose that were outside their ability to control and where they felt it was the Reddit administrators' job to intervene, even if the community as a whole was not in serious danger. These fell into two general categories. The first included ban evasion [4], where banned users make a series of new accounts each time they are banned and continue to cause trouble. Community moderators do not have the

capacity to tell for sure whether multiple accounts are run by the same user, so they contact the admins to handle the problem.

> *"A user gets banned, then gets on a tirade about why they shouldn't be banned and you mute them. And then they create 50 accounts to message you and that's ban evasion so you go to the admins and tell them to ban them."* – R9

Another difficult to handle issue for moderators is called brigading, when users from one subreddit decide to come together to vote on content in another subreddit to distort its popularity, to create a large amount of disruptive content, or to harass users and/or moderators. Brigading is strictly against Reddit's Content Policy, but it does occur. In these cases, moderators may contact Reddit administrators for help both in dealing with the immediate wave of problematic behaviors and in handling the follow-up with the community that committed the offense.

A final case where moderators seek support from administrators includes situations that moderators could probably control but that they feel more comfortable delegating to platforms. For example, moderators mentioned contacting the administrators when extreme content was posted that might violate laws, like attempted abuse of underage users or non-consensually posting sensitive imagery. While moderators could simply delete the content in question, the serious legal nature of the offense meant that they often preferred to pass responsibility to the Reddit admins.

> *"With the child predator stuff, we report it and let the admins do what they have to do. They don't give us any feedback, they just let us know that the report has been taken into consideration."* – R14

Broadly, volunteer moderators on Reddit feel that their role is to manage their communities, and they feel that they should have general autonomy to do so as they see fit, at least as long as their community adheres to basic standards of decency. They are willing

to sacrifice this autonomy in cases where they cannot handle something happening in the community, and when they feel that something is happening that is more serious than they want to take responsibility for. This combination of independence and inter-reliance could be termed "negotiated social autonomy".

Development of a multi-stakeholder process of deliberation:

A multi-stakeholder process of deliberation requires, by definition, multiple stakeholders to be aware that decisions are being made. As noted in prior work [19], moderators are generally unsure about how their specific communities are being evaluated by Reddit administrators. Nearly every Reddit moderator interviewee expressed uncertainty about how often, if at all, Reddit reviewed their community for violations of the Content Policy. Several moderators said they suspected that Reddit administrators didn't even know their community existed.

> *"My feeling is that [the admins] pay very little attention to my subreddit. I think it's just too small for them. It's less than 500k users, which by Reddit standards is small. I think it has a reputation for being capably managed so I think that given its reputation and its mods are actual experts and actual [practitioners of the discipline that is the topic of the subreddit], half the team has a PhD in it, I think they're willing to let us run our own affairs."* – R8

Though the processes for internal Reddit review of subreddits are not transparent, the public outcomes of these internal deliberations are the announcement of decisions to ban or quarantine specific subreddits. Moderators and regular users are frequently blindsided, with communities being removed that they were not aware were under investigation. When this occurs, users are frequently quick to critique the transparency of this process and to point out what they perceive as inconsistencies in the standards for punishment of a subreddit.

> *"I think the number one problem that Reddit has had with the quarantining and banning of subs is that they have not been consistent with their application of*

> *their own rules. I think the answer is you have a consistent set of rules for what is acceptable and what's not acceptable and you apply them consistently.* – R6 reinterview

Note that, in this case, and more broadly across Reddit, moderators are not directly asking to participate in the processes of deliberation, but rather to have a window into how decisions are made. The third step, the development of a multi-stakeholder process of deliberation, goes beyond just transparency, but greater transparency may perhaps be a necessary precursor to a fully participatory deliberative process.

Perhaps the most clear example of Reddit administrators' struggles with the cyberlibertarian values of transparency and openness was the years-long back-and-forth between /r/The_Donald (known colloquially as T_D), the primary subreddit dedicated to supporting Donald Trump. The subreddit, created for his 2016 presidential campaign, frequently hosted extreme far-right content that expressed hate and even calls for violence against minorities, particularly including transgender people. For years after the creation of T_D, users from across Reddit called for its banning, but Reddit administrators took a decidedly slower approach.

> *We [moderators] thought that was professionally irresponsible on their part when the_donald came up [...] [it] was a gathering place of the alt right, and they would discuss making organized attacks against other subreddits and so on, but [Reddit] couldn't ban it because it was ostensibly about an electoral candidate. I thought maybe they were waiting until after the election to ban it, when it seemed that Donald Trump would not be elected president, which as we know didn't quite happen that way. But I would say that the feeling on one end is Reddit admins spent a lot of time not doing enough, but also that at some point it got to the point where they actually couldn't do anything about it.* – R8 Reinterview

Reddit administrators' first strategy for engaging with T_D was to ban the most egre-

gious offenders within the community, which was unsuccessful in resolving the problems. Next, in late 2016, they took away privileges from T_D that had allowed members to artificially boost posts' visibility to reach a broader array of users outside the community.[8] In April of 2018, in response to a user question about why T_D had not been banned, Reddit CEO Steve Huffman (known as "spez" on Reddit) wrote that, while he found the community's explicit racism to be repugnant, it did not violate the sitewide rules. He said that "I believe the best defense against racism and other repugnant views, both on Reddit and in the world, is instead of trying to control what people can and cannot say through rules, is to repudiate these views in a free conversation, and empower our communities to do so on Reddit."[9] Over the next year, Reddit administrators attempted to work with T_D moderators to address the prevalence of rule violations, but, after more than three years of hoping that T_D would self-reform, Reddit administrators finally took serious action in June 2019 by "quarantining" T_D.[10] This step was followed by more serious pressure on the moderators of T_D, and the community eventually left the platform to form their own site. Speaking a year later in June 2020, Huffman wrote that he wished he had taken action sooner –

> *As we looked to our policies, "Breaking Reddit" was not a sufficient explanation for actioning a political subreddit, and I fear we let being technically correct get in the way of doing the right thing. Clearly, we should have quarantined it sooner.*
>
> ...
>
> *I admit we spent too much time with [T_D's] moderation teams that claimed to be doing their best while large numbers of users upvoted content that clearly broke our policies, which made it clear the issues were not of moderation, but of the*

[8] http://web.archive.org/web/20200502030704/https://www.theverge.com/2016/11/30/13797712/reddit-trump-the-donald-ban

[9] https://web.archive.org/save/https://old.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/dx5go62/

[10] Quarantining a subreddit means that users must click through a warning page to access it and also that posts to the subreddit are not indexed in searches and have much less visibility across the platform.

community culture.[11]

While banning and quarantining subreddits has traditionally been under the purview of Reddit administrators, in the case of T_D the subreddit was causing harm over a long period of time to the broader set of communities on Reddit and the administrators failed to stop it. It isn't possible to know how this situation might have played out if a true multi-stakeholder process for deliberation been in place, but Reddit users would at minimum have been able to make their arguments for action against T_D in a more formal environment than in the comments on posts made by administrators.

Reddit has already begun experimenting with a basic form of a multi-stakeholder process of deliberation via communication with "advisory councils" of moderators[12], a process that began in early to mid 2019, though this form of engagement is described more as an active way of giving feedback than as a process of public deliberation. As discussed below, the specifics of how these councils will function are not yet clearly outlined, but they will most likely have some ability to give feedback on proposed tools and changes to policies, a promising move toward tackling the fourth step.

Translation of outcomes of deliberations into 'terms and technologies':

The final step in the process of achieving a state of cooperative responsibility is to take the results of multi-stakeholder deliberation and translate them into things like policies, community guidelines, and tools. On April 29th, 2020, Reddit introduced a "Start Chatting" feature that would appear as a separate chatroom on each subreddit, ostensibly to encourage more social connections in a world where COVID-19 had forced social distancing.[13] These chats would be moderated by Reddit's Safety Team, rather than moderators of the host

[11] https://web.archive.org/web/20200605213844/https://old.reddit.com/r/announcements/comments/gxas21/upcoming_changes_to_our_content_policy_our_board/ft07637/

[12] https://old.reddit.com/r/modnews/comments/gw5dj5/remember_the_human_an_update_on_our_commitments/

[13] http://web.archive.org/web/20200504110128/https://www.reddit.com/r/blog/comments/gacdqy/new_start_chatting_feature_on_reddit/

subreddits. Reddit moderators quickly expressed extreme dissatisfaction with the change, with one post accumulating nearly one thousand comments almost exclusively opposed to the new feature, arguing that the new feature would sabotage the efforts of moderators who have spent years curating high-quality subreddits but would now be unable to apply this same care to a directly connected space. One moderator argued that Reddit has consistently focused on only three stakeholders – admins, general users, and advertisers, without considering the needs of moderators.[14] Within 24 hours, Reddit had completely rolled back the feature and apologized, stating that they had made a mistake – "We will not roll the feature out within your community again without having a way for you to opt out, and will provide you with ample notice and regular updates going forward."[15]

An interviewee described a related but inverted situation where his subreddit built a custom tool to meet their needs, but the tool was taken down by Reddit administrators:

> *"They [the admins] killed one of our bots called [botname]. On one hand we were annoyed because we were very up front and public about the launch, the mission, the purpose of the bots. The community was generally ok with our bots, but the admins didn't like the idea of [what they saw as] astro turfing. Then when we asked for it to be reinstated they didn't get back to us. Our bot was killed by one of their bots. Its nothing personal on one hand, but it seemed like to me that [subreddit] was not large or important enough to pay attention to our concerns."*
> – R8

These examples are not an outcome of a realized state of cooperative responsibility. In fact, they show evidence that supports the above assertions that the essential public values have not been collectively defined; Reddit administrators appear to misunderstand the values of the moderators. However, these examples show a starting point from which this type of

[14] http://web.archive.org/web/20200507200030/https://www.reddit.com/r/ModSupport/comments/gafm52/mods_must_have_the_ability_to_opt_out_of_start/fp124ks

[15] http://web.archive.org/web/20200507195210/https://www.reddit.com/r/ModSupport/comments/gafm52/mods_must_have_the_ability_to_opt_out_of_start/fp0r557/

process could be built. Reddit does as a platform have an ethos of open discussion and "rational" discourse, and this ethos could be channeled into the development of a more formal process of public deliberation and exchange. Ultimately, completing the steps that Helberger, Pierson and Poell proposed will require Reddit to create a process whereby a broader set of stakeholders (beyond just users and moderators) can actively participate in the discussions that shape codes of conduct, community guidelines, and site features. The advisory councils are one initial step toward this goal, but the role of these councils must be made clear and must be collectively agreed upon, and this role must expand beyond simply giving feedback. Moreover, stakeholders like governments and perhaps even advertisers (per the specific description of Helberger, Pierson, and Poell's second step) must be brought into the conversation to discuss their roles in the Reddit ecosystem.

3.4.2 Twitch and Cooperative Responsibility

Twitch is a video livestreaming platform that focuses primarily on social gameplay.[16] A "streamer" will livestream themselves playing a game, and viewers watch this video stream and chat in a chatroom that is associated with the stream. Streamers frequently directly interact with users by responding orally on the livestream to users' comments in the chatroom. Twitch, as a company, is similar to many other social platforms in that it relies on users to create (livestream) content, and it monetizes this content via advertisements shown on the streams and by taking a cut of paid subscriptions and "tips" to streamers. Twitch was purchased by Amazon in mid-2014, and has been integrated with Amazon's broader business plan both via use of Amazon Web Services to support the site's core streaming and chat functionalities and via the integration of Amazon's Prime service with Twitch in the form of "Twitch Prime", where Amazon Prime members receive additional benefits on Twitch. Twitch reserves the final say in determining what content appears on the site. However, the vast majority of in-the-moment moderation is done by volunteer user moderators who

[16]Though gameplay is the most popular category on Twitch, there are other categories including creation of music or art or "Just Chatting".

moderate virtually all of Twitch's hundreds of thousands of active stream communities.

Twitch exists in a complex cultural space; conflicts within gaming culture along the lines of race and gender (e.g., extensive targeted harassment during the Gamergate movement [13]) have spilled over on to Twitch in some cases, and the site frequently faces questions about the place of women in gaming (as I discuss below). Their communication with users is thus situated within these cultural conflicts, but is further complicated by the nature of the commercial structure of the site; unlike on platforms like Facebook and Reddit where monetization happens more behind-the-scenes, Twitch actually has direct business relationships with many streamers, who each generate revenue directly for the company via cultivating a "fan base" and attracting tips and subscriptions. As these streamers are all, by default, moderators of their own communities, Twitch does have direct communication channels with at least the monetized subset of streamers, and has at least some incentive to listen to their moderation concerns. The following sections analyze these relationships in depth.

Collective definition of essential public values:

The core values of Twitch are a complicated combination of Silicon Valley tech culture, classic gaming culture, and the influencer culture that has developed as a result of the rise of social media celebrities (e.g., Instagram stars). The product that Twitch provides is community-centered performance, where streamers create livestreamed video content and viewers engage with them and each other [18]. Like Reddit, Twitch moderates primarily at the community level. Rather than moderating individual messages sent in channel chatrooms, Twitch provides tools for volunteer moderators to do this and instead focuses on moderating streamers and their communities. Thus, while Twitch's community guidelines[17] do include rules for regular users, many of the rules are focused on acceptable conduct for streamers.

Across my interviews, streamers and moderators frequently used the same language to describe their main rules.

[17] https://web.archive.org/web/20200615191322/https://www.twitch.tv/p/legal/community-guidelines/

> *"So, I mean I think most streamers, including myself, have this sort of mentality that the first rule is... don't be a dick"* – T19

This rule, used almost verbatim across a wide variety of channels, translates roughly to "act like a reasonable, mature person", which in turn relies on a community-based understanding of what "reasonable" and "mature" mean.

Other common rules across different communities include "Don't be racist", "Don't be sexist", "Don't be homophobic" and occasionally "Don't be transphobic" or "ableist".

> *"[My main rules are] the obvious ones like racism, homophobia, any kinds of bigotry really aren't tolerated"* – T20

> *"So the only rules that I have are essentially just no racism, no sexism, no homophobic remarks. So outside of that I don't care what you say, as long as you're not being too controversial I guess."* – T7

These types of rules nominally imply a general agreement over values of tolerance on the platform, but at the heart of Twitch's culture is a clash over identities that mirrors the broader clash in the gaming industry.

Among the most controversial value-related issues on Twitch has been the way women are presented and present themselves on the platform. Female streamers are a relatively small minority on the platform, and many regularly face abuse and harassment because of their gender. As of late 2017 and early 2018, a number of prominent female streamers had gained popularity on Twitch in part by presenting themselves in a non-nude but very sexualized way and performing on stream in a way that foregrounded their bodies. In response, in February of 2018, Twitch released a new set of detailed guidelines for self-presentation on stream,[18] which was largely focused on specifying exactly what types of clothing and behavior was permissible for these types of female streamers (See Fig 3-2 for a subset of these guidelines).

[18] https://web.archive.org/web/20200626111412/https://www.twitch.tv/p/legal/community-guidelines/sexualcontent/

Sexually Suggestive Content

To maintain the health of our community and promote content that is appropriate for a diverse audience, sexually suggestive content is prohibited on Twitch. Evaluations on the sexual suggestiveness of a behavior or activity are independent of user attire and are instead based on the overall surrounding framing and context. This policy also applies to embedded media, augmented reality, creative broadcasts, and channel content—such as banners, profile images, emotes, and panels—that are focused on provocative images or video.

Content that is considered to be sexually suggestive includes, but is not limited to:

- Content or camera focus on breasts, buttocks, or pelvic region, including poses that deliberately highlight these elements
- Groping or explicit gestures directed towards breasts, buttocks, or genitals
- Fetishizing behavior or activity, such as focusing on body parts for sexual gratification or erotic role play
- Simulated sex acts or sexual stimulation
- Using or featuring sex toys in contexts unrelated to sexual education
- Erotic dances, such as those involving stripping or flashing
- Pole dances or acrobatics with sexually suggestive framing
- Posting, displaying, or sharing erotica, including detailed descriptions of sex acts or pornography

Figure 3-2: Prohibited Sexually Suggestive Content on Twitch, Posted Feb 7th 2018.

Interviewees had mixed opinions of these new rules, with some opposing them on the basis of the idea that women's bodies shouldn't be policed by platforms and others supporting them because they felt that Twitch needed to avoid becoming a "camgirl" platform.

> *"I have mixed feelings about the appropriate attire bullshit. I think on one hand, um, it's aimed at women and uh, it is policing women. And how they look. And I think that's Kinda bullshit. [...] And on the other hand, like maybe we should have like a better gating for age appropriateness for twitch streams."* – T2 Reinterview

Gerrard and Thornham write that community guidelines are one major place where platforms express gendered value systems [7], and Twitch's guidelines are unique in that they establish gender-specific norms for behavior of women (and men), not just in images or video, but in real time interactions.

Gendered behavior is among the most visible controversial issues on Twitch, and it is likely one that will continue to be controversial absent a large cultural shift, which can

only happen as a result of large-scale engagement with the topics in question. However, many other value questions ranging from racist harassment and transphobic attacks were mentioned by moderators.

> *"There is still a lot of more specific things that I think they'd need to go after. Particularly specific forms of harassment like mis-gendering, deadnaming, doxing. [...] I think really just the general thing would be [Twitch] need to put some time into carving out very specific direct and clear policies about what is and isn't harassment. Um, cause just saying don't harass people doesn't really help anybody."* – T16

Unlike on Reddit, there are no places for open discussion between administrators and users about these policies. While there may be behind-the-scenes conversations between established streamers and their contacts at Twitch, and many users and streamers regularly voice their opinions, there is no forum on Twitch where the administrators in charge of setting rules explain their values publicly via discussion back and forth with users. Thus, it is difficult to say that Twitch has made significant headway in collectively defining essential public values. In this sense, though, it may be fair to say that Twitch has a greater challenge than Reddit. While the latter was a platform created *from* a value set into a space with many similar text-based platforms, Twitch was created to host a relatively new type of content that could be produced by people with a very wide range of values.

Stakeholders' acceptance that they have a role:

In the Frequently Asked Questions section of the Community Guidelines,[19] Twitch notes that:

> "Creators are role models and leaders of the communities they create or foster around them. Creators should consider the consequences of their statements and

[19] https://web.archive.org/web/20200624203820/https://www.twitch.tv/p/legal/community-guidelines/faq/

actions of their audiences; we ask that you make a good faith effort to quell any efforts from those in your community to harass others."

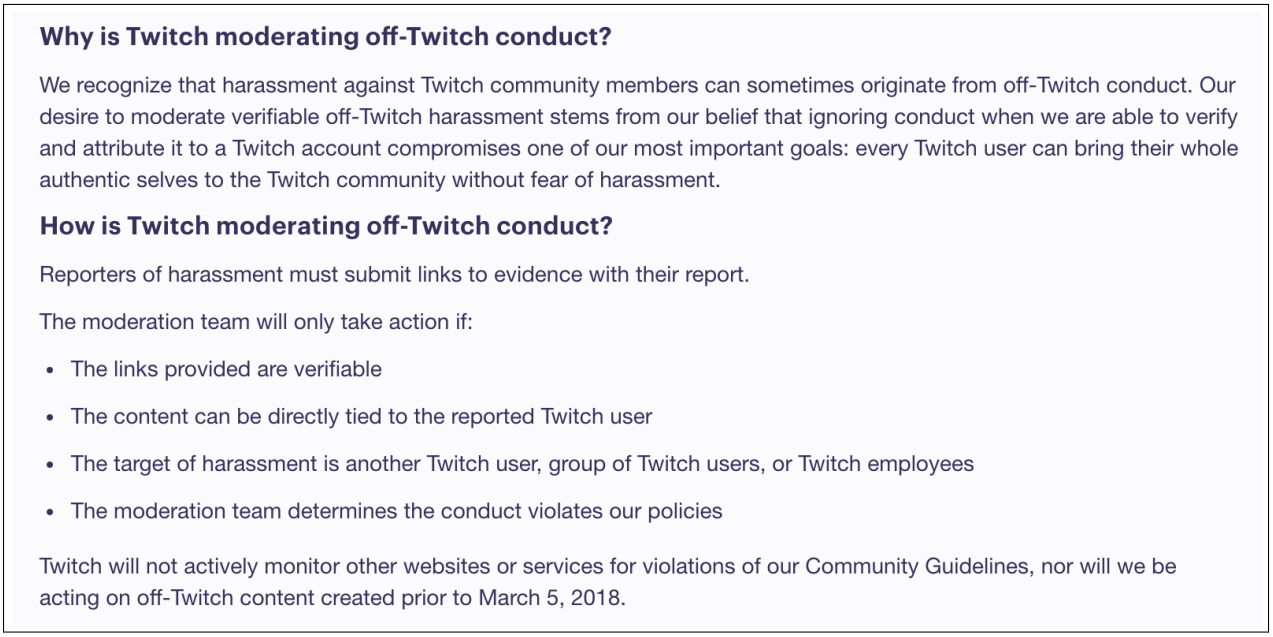
Like Reddit, Twitch delegates the majority of the moderation within communities to streamers and their moderator teams, instead focusing on moderating streamers. Streamers accept they are responsible for the content of their stream and to some extent what their community does.

> *"I feel at like, at Twitch's level, it's more about managing broadcasters, right, like I manage my community and I expect every other Twitch streamer to do the same. [...] I think that Twitch's job is to manage broadcasters, to control that space, so if a broadcaster is being disruptive, and harmful to the community at large, it's Twitch's job to manage them."* – T16

One consequence of Twitch's format is that it leads to the formation of strong communities [9] and also celebrity streamers; whereas subreddits are based around topics and may have some fairly visible power-users, Twitch channels are literally based around one person who is in most cases literally visible. The classic power law distribution of content creation, exists on Twitch as on many other platforms, so a relatively small number of streamers account for the vast majority of viewers (and likely also subscriptions and ad revenue) on the site [17]. When a moderation action is taken against a visible streamer, it can become a heated topic of debate and often results in an outcry from the streamer's "fans". However, unlike on Reddit, Twitch maintains a strict policy of not commenting on punishments given to streamers or breaches of the community guidelines, so users are often left to speculate about the true reasons for a punishment (absent a convincing explanation from the streamer in question).

Twitch is rare among major social platforms in that it actively reserves the right to make moderation decisions based on actions that happen off the platform, e.g., on Twitter or even in real life, so long as they directly impact the Twitch experience. While it may seem like an

overreach for Twitch to include this in its self-defined role, it fits well with Twitch's goal to be a social hub; many visible streamers on Twitch have an active social media presence on other platforms, and this type of intervention (ideally) discourages streamers from simply harassing each other on a different platform when they are forbidden to do so on Twitch. The Twitch experience, unlike the Reddit experience or the Facebook experience, is deeply and inseparably linked to social interactions on Twitter, Discord, and sometimes YouTube and Instagram.



Why is Twitch moderating off-Twitch conduct?

We recognize that harassment against Twitch community members can sometimes originate from off-Twitch conduct. Our desire to moderate verifiable off-Twitch harassment stems from our belief that ignoring conduct when we are able to verify and attribute it to a Twitch account compromises one of our most important goals: every Twitch user can bring their whole authentic selves to the Twitch community without fear of harassment.

How is Twitch moderating off-Twitch conduct?

Reporters of harassment must submit links to evidence with their report.

The moderation team will only take action if:

- The links provided are verifiable
- The content can be directly tied to the reported Twitch user
- The target of harassment is another Twitch user, group of Twitch users, or Twitch employees
- The moderation team determines the conduct violates our policies

Twitch will not actively monitor other websites or services for violations of our Community Guidelines, nor will we be acting on off-Twitch content created prior to March 5, 2018.

Figure 3-3: FAQ about how Twitch takes off-platform content into consideration.

The main, role-based point of conflict between users and Twitch administrators is disagreement over when and how Twitch should step in to address conflict and respond to accused harassment. As noted above, many Twitch users believe Twitch has a responsibility to take a stronger stance against sexualized behavior from female streamers. Other users believe that TWitch has not adequately responded to cases of harassment against minority users on the platform. This issue, which exists at the intersection of roles and values, has thus far eluded resolution. As on Reddit, interviewees were primarily concerned with issues of consistency and transparency, and whether rules were enforced at all.

> *"Honestly, it doesn't matter because they don't enforce it. They threw up some words on it in a Twitch announcement to try and incur good favor from the queer community of twitch and then have done literally nothing with those policies. Since I've reported, I've reported multiple major and minor streamers directly to the Twitch staff I know, and nothing happens."* – T19 reinterview

Development of a multi-stakeholder process of deliberation:

Twitch administrators do not engage in the same open modes of discussion that Reddit administrators use. While, on Reddit, these discussions are prominent, easily accessible, and fully archived, there is no comprehensive, public, archived record of conversations between users and administrators on Twitch. Part of this is likely due to the structural differences between the platforms; Twitch streamers who have reached any level of popularity have business contracts with Twitch and receive a proportion of ad revenue and user subscription fees, while it is strictly against the ethos of Reddit (and its rules) for subreddit moderators to profit off of the communities they run. This makes sense – on a subreddit, a wide variety of users contribute content intermittently, while on a Twitch channel the streamer is the primary content creator who may spend an amount of time creating content that parallels the time spent on a traditional full-time job. In these cases, needs are communicated largely via backchannels, e.g., partner managers, but more broadly via personal networks and personal connections with Twitch employees.

> *"I think back then [when I just became a partner] the issue was that I didn't know the systems or like the lines of communication as well as I do now and now I know where to go and who to talk to when I need to do something or ask something or what have you. Back then I'd become a partner and then sort of just been left to my own devices, you know? I know a couple of people more personally now, so it's a lot easier to interact in that way."* – T19 reinterview

Twitch's major contribution toward including users in deliberation was the introduction of

the Twitch Safety Advisory Council in May 2020, which, though it occurred after interviews concluded, is relevant to mention. The council was initially comprised of eight members, four of whom were accomplished streamers, two were advisors from nonprofit advocacy groups, and two were academics doing research in relevant areas.[20] Per a blog post by Twitch CEO Emmett Shear,[21] the purpose of the Council is to "advise, offer perspective, and participate in discussions with our internal teams pertaining to the work we do to help keep our community safe and healthy." Unlike Facebook's Oversight Board, "Council members will not make moderation decisions, nor will they have access to any details on specific moderation cases. They are not Twitch employees, and they do not speak on Twitch's behalf." Though this council, both through its academic expertise and diversity of represented identities, may help Twitch write more inclusive policies that are sensitive to the needs of underrepresented users, this council does not by itself provide a space for deliberation between stakeholders. Only a very select group of users are permitted to give opinions, and none have formal decision-making authority.

Translation of outcomes of deliberations into 'terms and technologies':

Given the lack of any processes for deliberation, it is difficult to say that outcomes of deliberation have been translated into 'terms and technologies'. It is plausible that future iterations of the community guidelines may be influenced by feedback solicited from the advisory council, and it is likely that moderation tools and platform features will be influenced by consulting with power-users and outside experts, but these are not the results of any sort of public deliberation. In order to achieve this step, Twitch will need to identify a method for soliciting public feedback as part of a deliberative process.

[20] https://web.archive.org/web/20200626095228/https://blog.twitch.tv/en/2020/05/14/introducing-the-twitch-safety-advisory-council/

[21] https://web.archive.org/web/20200626102614/https://blog.twitch.tv/en/2020/05/19/a-note-from-emmett-about-the-safety-advisory-council/

3.4.3 Facebook Groups and Cooperative Responsibility

Though significantly less discussed in public discourse and research literature than the primary Facebook network functionality, the groups portion of Facebook contained more than 200 million groups as of mid-2018.[22] Each of these groups has at least one moderator and many have multiple [19]. Facebook groups span a broad spectrum of topic areas, with some devoted to mutual interests, others to buying and selling, others to local communities, and some just allowing extended groups of friends a shared place to post.

The concept of cooperative responsibility is perhaps most difficult to apply to Facebook groups because user-moderated groups exist alongside and in connection with the non-group, network portion of Facebook, which is moderated solely by the company. The platform as a whole was solely network-structured in its first iteration, with groups arriving shortly after. On Facebook, users move back and forth between platform-governed spaces and user-governed spaces fluidly. Thus, cooperative responsibility must be seen as negotiations with different sets of stakeholders about related spaces, but where the stakeholder groups change and so do the responsibilities depending on the space. In a process for achieving cooperative responsibility, the stakeholders in the network space would be Facebook and users, whereas in the group space the stakeholders would be Facebook, moderators, and non-moderator users. I focus here on the latter space, but note the connections between the two spaces as appropriate.

Collective definition of essential public values:

The split between the network portion of Facebook and the groups portion is reflected in users' perceptions of Facebook's values. While users have some general idea of what types of content are moderated on the network, and most of Facebook's press releases deal with the network portion rather than the groups portion, they have difficulty understanding what

[22] https://web.archive.org/web/20191208191626/https://singjupost.com/full-transcript-mark-zuckerberg-at-facebooks-f8-2018-developer-conference/?singlepage=1

Facebook's vision is for groups or what values Facebook holds.

> *"I mean, yeah, I just have no experience of [Facebook's] philosophy. So I'm drawing a blank. I guess [the best metaphor is] absentee parents or something."* – F7 reinterview

> *"The moderating style of Facebook, like as a platform, like what gets reported, what gets banned, it feels like if you report something they just run it through the algorithm again, right. With no human touch. Or when there is human touch, I feel like it's somebody who doesn't speak the language [who] has to translate it through like Google translate."* – F15 reinterview

While Reddit and Twitch both have very large international userbases,[23] Facebook is a truly global platform. Its sheer size and lack of a unifying focus (i.e., Reddit's free speech ethos or Twitch's livestreaming technology) makes it difficult to see how essential public values could be defined. Scholars like David Kaye [11] have suggested that Human Rights Law could provide a foundation for defining essential public values, but Kaye imagines a scenario where platform values are less publicly defined and more adopted by Facebook administrators. Regardless of the challenges involved, it is safe to say that the moderators interviewed for this study did not feel that essential public values had been defined in any way shape or form.

Stakeholders' acceptance that they have a role:

The collective definition of roles, however, is a much simpler task on Facebook Groups. The role divisions used by Reddit and TWitch would both translate fairly well to Facebook groups; moderators[24] are responsible for the day-to-day operations of groups, while Facebook

[23]Per Statista, roughly half of Reddit users and roughly one fifth of Twitch users are from the United States.

[24]Facebook calls its lead moderators "admins" and second tier moderators "moderators", but for clarity I refer to both as moderators here.

is responsible for stepping in when whole groups become problematic. Facebook's response to a problematic group could take the Reddit approach, where Facebook admins attempt to work with the moderators to steer things back on track, or the Twitch approach, where temporary suspensions from the platform serve as notice that behavior has become problematic (though it is likely that Twitch also does some of the former). Facebook could also perhaps provide some services to Groups moderators akin to Reddit's support when things happen in the Groups that the moderators cannot handle or are uncomfortable dealing with.

In contrast to Reddit and Twitch's (mostly) two-tier models, many moderator interviewees complained about unexplained moderation interventions that Facebook had made *within* their groups without warning. Interviewees felt that these interventions, whether made by algorithms or humans, were not made with an understanding of the group's norms and culture.

> *"I've noticed a lot of groups have been getting shut down for the actions of maybe a member who posted like an inappropriate picture or link or a, a racial slur or something. And rather than just let the admins of a group deal with it, if somebody reports that comment, Facebook might just destroy the group altogether and remove it from the site. So we've had to be a little bit more careful about that. And for the most part, if there's something that really violates Facebook's terms of service rather than just our group's rules, we try to have members just, um, tag us or send us a message privately rather than report it to Facebook just because sometimes they'll go a little bit overboard."* – F5 reinterview

Broadly, moderators in Facebook Groups have a strong sense of what their role is but they do not understand what Facebook's role is supposed to be within the groups space. Models from Twitch and Reddit could inform development of a clearer role division in the Facebook Groups space.

Development of a multi-stakeholder process of deliberation:

Facebook has both formal and informal processes for connecting with users and getting feedback. The more informal process is simply reaching out to users when something unusual happens:

> *"There's a group that I was part of for a very long time that, when it fell apart, it actually got news attention. [...] It was given to me and a couple other people and like one of the original administrators [to deal with] and at this time you couldn't archive groups, you couldn't just shut them down entirely, so we turned off all the comments, we turned off all the new posting, and stopped accepting requests to the group. And then we ran a javascript thing that would just individually remove members one by one by one. And during that time, a Facebook employee reached out to us and just asked us like... this is kind of weird and a little unprecedented, why are you doing this? And we just explained to them that the moderation tools available at that time made it really difficult to run the group. [...] It was a really interesting conversation. We talked about some potential tools that administrators or moderators could use. And I think that my contributions affected things."* – F11

In this sense, Facebook happened upon a situation that gained them a deeper understanding of how groups work, but this was not a planned process. A more formal process moderator interviewees mentioned was the presence of Facebook-run Groups dedicated to user research on the platform.

> *"I was invited at one point to join some huge internal Facebook research thing where they had invited only admins of groups above a certain threshold. I initially joined the group, which was run by Facebook employees, but I chose not to participate because I kind of have a vendetta against Mark Zuckerberg. Not*

him personally, I should specify, the company. And I wasn't going to give them free market research into how they can control groups [...] so I was like I'm not going to go in there and give Mark Zuckerberg feedback on how he can improve his moderation tools because he's already kind of making my experience difficult."
– F13

The most formal method that Facebook uses to solicit feedback is focused on *external* feedback from experts (who may or may not be Facebook users) via its newly created Oversight Board [2], which was created to be an external arbiter and make judgments on controversial removal decisions. However, the Oversight Board was not initially created to have any oversight over activity within Groups; its scope was limited to the network part of Facebook. This could be interpreted in multiple ways. It could suggest that Groups are less of a focus at Facebook, but it could also be a sign of understanding the importance of Groups moderators in making decisions. If the Oversight Board had authority over Groups, would it be able to overrule volunteer moderators' decisions?

To date, there is no formal process on Facebook Groups that matches Helberger, Pierson, and Poell's conception of a deliberative process within the framework of cooperative responsibility, but each of the above processes contains one relevant element. The first, most informal process shows merit via communication directly with users and an attempt to understand specific circumstances before taking action. The second process creates a forum for discussion, albeit not a forum that tackles complex value or process questions. The third, most formal process recognizes the importance of deliberation about important issues, though its structure is far removed from the ideas of collective definition of essential public values, and the people engaged in the deliberative processes are academic and social elites rather than regular users.

Translation of outcomes of deliberations into 'terms and technologies':

Though the above represent initial steps toward a multi-stakeholder process of deliberation, such a process is not currently in full operation. In their interviews, moderators mentioned a number of issues that they would like to see resolved, which would likely be addressed if such a process were in place. Notably, these included issues related to moderator safety.

> *"The fact that there's no way to put on a mask when you're doing moderation especially in large groups, I think it's a huge failure in Facebook's moderation policies and API because I have been harassed literally just because of my status as a moderator or an admin in the group. And there's no way to hide that from anyone unless you block a person. And of course they could just make a sock[puppet] account and go around it. Due to my personal experiences, that's what I'd put at the top of where tools are lacking."* – F13

Another moderator mentioned issues surrounding respect for moderators from marginalized groups.

> *"A lot of our mods are people with marginalized identities. Some mods don't use their real names for either desire for semi anonymity on the internet or because they're trans mods. It's pretty common that if a mod who isn't using their real name is interacting with someone who's getting really angry, they'll get Zucked, which means they'll get reported for the real name policy. We had someone who only used their first name as their Facebook name... they got Zucked. So they end up having to interact with the Facebook review systems in that way."* – F14

A more thorough attempt by Facebook to work through the steps toward a process of cooperative responsibility would likely help Facebook and its users reach a common understanding on issues like these and many others, and this understanding could translate into 'terms and technologies' per Helberger, Pierson, and Poell's description.

3.5 Discussion and Implications

Each of the three platforms has approached the challenges inherent in the concept of cooperative responsibility differently. I conclude here first by comparing each platform's approach within each step and discussing their similarities and differences, and then by taking a step back and reconsidering how this work contributes to the theory of cooperative responsibility.

3.5.1 Platforms

Of the three platforms, Reddit has come closest to collective definition of essential public values, partly through its content policy and partly through Reddiquette, but also by how it brands itself. Reddit has always been a place for discussion of a wide variety of topics, and has positioned itself toward the extreme end of the spectrum of permissibility among social media platforms in that regard. Though Reddit's standards for what type of discussion to permit have evolved over time, the guiding philosophy and core values that the platform presents are similar to those present at its founding. Thus, it collectively defines values with its users in part because those users self-select on to Reddit. However, Reddit admins are also very active in engaging with users on value debates, and while this does not necessarily directly translate into an collectively agreed-upon set of values, it does make value-discourse central on the platform.

Twitch has neither a core foundation of values that sets it apart from other platforms nor an active arena for discussion between users and administrators. It is a platform for the presentation of a type of content, with values imposed in a largely top-down manner. As noted above, Twitch sits in a complex cultural space, trying to maintain a welcoming and inclusive platform in the gaming domain, but it is perhaps through this cultural angle that Twitch might reach a collective definition of values.

Facebook has none of the above grounding points for collective definition of essential public values. It has no strong founding values aside from the early Ivy League elitism that

it quickly cast aside in favor of growth-as-a-value. It has no specific cultural niche that it fills, instead aiming to touch all cultures and interests at a surface level. It also has no forum for public discussion between users and Facebook employees in which values might be defined. It is possible that subdivisions of Facebook might be better equipped to handle these processes – could values be better defined if the Facebook groups ecosystem were divided into a number of cultural sub-units, each with its own governance systems?

Despite their differences in terms of value definition, each of the three platforms maintains a similar role division; on each, the platform delegates most of the day-to-day work of moderation to volunteer moderators in each community, and focuses on managing the space as a whole. When a group or community proves beyond the ability of moderators to handle, the platform intervenes. There are two ways in which these role divisions vary. First, the "remediation" processes differ across the three platforms. On Reddit, a relatively clear process is outlined; Reddit admins will work with moderators of the communities in question until either the community is under control or it becomes clear that the moderators are uninterested in cooperating. On Twitch, the process is less public; as Twitch's relationship with streamers is a business relationship, communications about acceptable conduct are kept confidential. However, as noted above, warnings are sometimes given and streamers can be given multiple smaller punishments before being permanently banned from the platform. It is unclear whether any type of remediation process exists for Facebook groups. None of the interviewees was aware of such a process, and many had heard stories of groups being removed with no warning, so if such a process exists it is not always present. The second way role divisions vary is in terms of how active the platform is in intervening within groups. Twitch virtually never moderates chat messages within a stream. Reddit will occasionally remove messages within subreddits, but only for a specific set of predefined reasons and may also notify the moderation team. Facebook seems to be active in moderating posts within groups, but it is unclear to group moderators what types of posts are being moderated, how they are being assessed, and when an intervention has been made. This makes the division of

roles between moderators and Facebook admins far less clear than on the other platforms, as Facebook admins can at any point take action within the groups that moderators expected to be running primarily on their own.

Beyond Reddit's open forums for discussion between users and administrators, which neither other platform has attempted, the primary process for engaging external parties in deliberation is through each platform's advisory or oversight board or council. Though there are many factors that might be considered in evaluating these bodies, I focus here on two: their composition, and their authority.

As it will be the standard against which each other board is measured, at least in public conversations, I will begin with Facebook's Oversight Board. The Oversight Board's first twenty members, announced in early May 2020,[25] are drawn heavily from legal backgrounds. Fourteen are current or former lawyers or legal scholars. The remaining six include three journalists or scholars of journalism, one communications professor, one political scientist, and one former government official. This composition matches the extensive array of legal metaphors used to describe the board, including "Facebook's Supreme Court" [2]. The "cases" that the board initially reviews will be users' appeals regarding pieces of their content that have been taken down. Groups and Pages are not included in the Board's initial purview, but this may change in the future. The expected duration for a typical "case" will be 90 days, with the possibility for expedited cases completed within 30 days[26]. Facebook may also specifically request that the Board review certain pieces of content [12]. While the Board's decision on any particular piece of content is binding, Facebook is not obligated to take the same action on identical pieces of content in the future [12]. Realistically speaking, the Facebook Oversight Board is not designed to be an active arbiter of speech. The internet moves quickly; a piece of content that was removed but reinstated 90 days later (or even 30) has lost its potential for impact. The same is true, if not more so, if the board were given

[25] https://web.archive.org/web/20200529153029/https://www.politico.com/news/2020/05/06/facebook-global-oversight-board-picks-240150

[26] https://web.archive.org/web/20200603024045/https://about.fb.com/news/2020/01/facebooks-oversight-board/

the authority to determine whether to take down particularly controversial pieces of content. By the time the board made its decision 30-90 days later, the piece of content would have already reached its audience. Thus, the Facebook Oversight Board can be seen primarily as a mechanism for facilitating conversations about content moderation on Facebook and nudging Facebook to be more transparent in certain circumstances.

Twitch's Safety Advisory Council, announced May 14th 2020,[27] is designed, per Twitch CEO Emmett Shear, to "advise, offer perspective, and participate in discussions with our internal teams pertaining to the work we do to help keep our community safe and healthy."[28] In this sense, this council mirrors many other similar councils that a variety of other social media platforms have made use of, bringing in external experts to advise on policy. As noted above, the Safety Advisory Council does not have decision-making power over specific cases. The primary notable difference between this approach and Facebook's, however, is the composition of the council; where Facebook's Oversight Board was dominated by lawyers and legal scholars, Twitch's Safety Advisory Council's eight members include four prominent users of the site, one sociology professor, one criminology professor, one anti-bullying activist, and one leader of an advocacy group for free expression, working from a legal background.[29] The incorporation of active users of the site, particularly including users who are from marginalized groups, helps represent real users' experiences in the advising process.

Reddit has gone even further with this user-centered advisory model. As noted above, Reddit has been experimenting with "advisory councils" since mid-2019, which are groups of moderators who are consulted quarterly for their opinions on new tools and policies.[30] Alex

[27] https://web.archive.org/web/20200626095228/https://blog.twitch.tv/en/2020/05/14/introducing-the-twitch-safety-advisory-council/

[28] https://web.archive.org/web/20200626102614/https://blog.twitch.tv/en/2020/05/19/a-note-from-emmett-about-the-safety-advisory-council/

[29] Facebook's Oversight Board is unique in how lawyer-heavy its membership is; TikTok's oversight board is composed of seven members including two legal scholars, a tech ethicist, an academic social worker focusing on children's safety, an academic image and video processing expert, a political scientist, and a technology policy expert (https://web.archive.org/web/20200518021836/https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council)

[30] https://old.reddit.com/r/modnews/comments/gw5dj5/remember_the_human_an_update_on_our_commitments/

Le, Reddit VP for Product, Design, and Community, wrote that "These council meetings have improved the visibility of moderator issues internally within the company."[31] Like with Twitch's Safety Advisory Council, these moderator councils do not appear to have the ability to adjudicate "cases" (beyond the content that they already moderate on their own subreddits). On June 29, 2020, Le released summary notes from a recent call between 5 Reddit executives, 20 Reddit staff, and more than 25 of these moderators, in which moderators gave feedback on proposed policy changes and also highlighted relevant current failures in the tools space.[32] Though the notes do not show how the policies were specifically impacted by feedback from the moderators on the call, they do clearly show that numerous high-level Reddit employees are directly soliciting feedback from users about policies in a detailed, thorough manner. It is unclear whether the notes from these calls will continue to be made public in the future, but doing so could help bring more attention to the nuance of these moderation issues on Reddit and more broadly.

One point to note about all of the above boards and councils is that none of them contain regular, average users. While Twitch's Safety Advisory Council and Reddit's moderator councils both include users, these users are already users with authority (i.e., streamers, and moderators of significant subreddits). Though the inclusion of these users in the conversation is a worthwhile starting point, Helberger, Pierson, and Poell's concept of cooperative responsibility does not propose a division of labor agreed upon by platforms and *powerful* users; it suggests that all users should have a voice in shaping this division of labor [10, p. 10].

3.5.2 Implications for Cooperative Responsibility Theory

This work also furthers the concept of cooperative responsibility in several ways. First, it identifies two challenges – one conceptual and one methodological. The first comes from the

[31] https://web.archive.org/web/20200604204005/https://www.reddit.com/r/modnews/comments/gw5dj5/remember_the_human_an_update_on_our_commitments/

[32] https://web.archive.org/web/20200630034929/https://www.reddit.com/r/modnews/comments/hi3nkr/the_mod_conversations_that_went_into_todays/

fact that Helberger, Pierson, and Poell do not explicitly provide an explanation of what it looks like to have achieved a state of cooperative responsibility; they present examples of situations where progression toward such a state could be valuable, and they identify four steps that could lead toward this state, but a clear description of the goal state is not present in their work [10]. The second challenge is in the difficulty of gathering complete data. In this chapter I have shown examples of where platform administrators and community moderators communicate (or fail to communicate), and I have detailed the ways in which moderators experience and react to these successes and failures. This data does not include interviews with platform administrators, and as such I can only make inferences about administrators' motivations based on public statements and documented incidents. It will be a challenge for researchers exploring the concept of cooperative responsibility in any business context to get access to publish analysis of interviews with platform administrators discussing whether they are considering what could be a significant change to the platform's structure and perhaps even business model.

Next, I have identified two advantages to this theoretical approach. First, as this work has shown, cooperative responsibility can theorize systems with many stakeholder groups that are internally diverse. Helberger, Pierson, and Poell [10] make references to diverse stakeholders, but focus less on diversity within stakeholder groups. In the case of platforms that make use of community-driven moderation, communities have widely varied goals and values, but the deliberative processes inherent in the concept of cooperative responsibility show promise for being able to negotiate these differences, albeit only if the deliberative processes are well-designed. The second advantage is that cooperative responsibility foregrounds the experiences of users. As I explored in depth in the literature review chapter, the dominant paradigm in academic and public discourse has been the one that focuses on the role of platforms in moderation. The core of the theory of cooperative responsibility is the idea that users are equally as important to focus on as platforms; it suggests a paradigm where platforms and users are studied in parallel.

While each of these three platforms has taken some actions that correspond to the first three steps of Helberger, Pierson, and Poell's vision for a path toward a state of cooperative responsibility, none of the three have yet achieved a state where the fourth step can be said to be occurring; there are not yet deliberations that are sophisticated enough to produce 'terms and technologies' that represent agreement between many involved stakeholders. With this chapter I hope to bring greater attention to the framework of cooperative responsibility and to demonstrate its applicability to the context of content moderation on platforms that rely heavily on user labor. I hope that, through highlighting some of the actions these platforms have taken so far this work can help push the conversation over whether and how a more complete state of cooperative responsibility might be achieved.

Bibliography

[1] 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. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):31:1–31:22, December 2017.

[2] Evelyn Douek. Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility. *N.C. J.L. & Tech*, 21:1–78, 2019.

[3] Andrea Forte, Vanesa Larco, and Amy Bruckman. Decentralization in Wikipedia Governance. *Journal of Management Information Systems*, 26(1):49–72, 2009.

[4] Eric J. Friedman and Paul Resnick. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2):173–199, 2001.

[5] Clifford Geertz. *The interpretation of cultures*. Basic books, New York, NY, USA, 1973.

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

[7] Ysabel Gerrard and Helen Thornham. Content moderation: Social media's sexist assemblages. *New Media & Society*, July 2020.

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

[9] William A. Hamilton, Oliver Garretson, and Andruid Kerne. Streaming on twitch: Fostering participatory communities of play within live mixed media. In *Proceedings*

of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, page 1315–1324, New York, NY, USA, 2014. Association for Computing Machinery.

[10] Natali Helberger, Jo Pierson, and Thomas Poell. Governing online platforms: From contested to cooperative responsibility. *Information Society*, 2018.

[11] David Kaye. *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports, New York, NY, USA, 2019.

[12] Kate Klonick. The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *The Yale Law Journal*, 129(8):2418–2499, 2020.

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

[14] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[15] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK, 1990.

[16] Elinor Ostrom. Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3):137–158, September 2000.

[17] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The social roles of bots: Evaluating impact of bots on discussions in online communities. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):157:1–157:29, November 2018.

[18] Joseph Seering, Saiph Savage, Michael Eagle, Joshua Churchin, Rachel Moeller, Jeffrey P. Bigham, and Jessica Hammer. Audience participation games: Blurring the line between player and spectator. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, page 429–440, New York, NY, USA, 2017. Association for Computing Machinery.

[19] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[20] Dennis F Thompson. Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, 74(4):905–916, 1980.

[21] Fernanda B Viégas, Martin Wattenberg, and Matthew M McKeon. The Hidden Order of Wikipedia. In Douglas Schuler, editor, *Online Communities and Social Computing*, pages 445–454, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

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

4

Metaphors for Moderation

4.1 Introduction[1]

In a technical manual for the CommuniTree Bulletin Board System, founded in 1978 as one of the first public online communities, writer and engineer Dean Gengle discussed his philosophy for what he called a "Fairwitness":

> [T]he new social role of Fairwitness for [online] computer conferences combines elements of previous roles for the purposes of guiding the conference process, so it can be useful to its network of users. These roles are: conference moderator, editor, peacekeeper, promoter, guide, ombudsman, chairperson, host/ess, traffic manager, database organizer, pump-primer, and sometimes silent partner to the system operator(s). [8]

Gengle drew the term "Fairwitness" from Robert A. Heinlein's 1961 novel, *A Stranger in a Strange Land*, where "Fair Witnesses" are futuristic citizens who can perfectly recall any event without distorting it through their own prejudices. Gengle chose this term to underscore the importance of users setting aside their prejudices and tempering their emotions in handling sensitive matters in their online community.

[1] A modified version of this chapter, with Joseph Seering as first author and Geoff Kaufman and Stevie Chancellor as co-authors, has been accepted to *New Media & Society*.

Gengle's metaphor-laden articulation of the role of a Fairwitness was not, strictly speaking, the first description of what we might today call a "moderator" in an online community, though it is certainly one of the earliest philosophical explorations. In their 1978 book, *The Network Nation*, Hiltz and Turoff described moderation in the Electronic Information Exchange System, a system used by researchers to collaborate virtually within forums called "conferences". Just as offline academic conferences need moderators to oversee and facilitate conversation during sessions, they argued, so would online conferences [13, pp. 23–24].[2]

Metaphors are powerful tools for exploring moderation, explaining individuals' relationships, values, and roles in social situations [21]. Research from psychology shows how metaphors concretely shape our understanding of abstract concepts that influence problem solving and social coordination [25, 23, 24]. In this chapter I use analysis of metaphors to answer whether, as the technical landscape of moderation and online communities has evolved, the corresponding conceptual landscape of how moderators and community administrators view their roles has also changed.

Following in the footsteps of Gengle and of Hiltz and Turoff, modern academic literature has continued to apply metaphorical labels like "custodians" [12] and "governors" [20] to moderators in corporate, platform-moderated models of social media. Though recent literature has focused more on this platform-moderated model, volunteer community moderation is crucial to the content moderation ecosystem. In mid-2018, over 1.4 billion people used 200+ million Facebook Groups each month, each of which had, at minimum, one volunteer moderator and often multiple.[3] Reddit reports more than 430 million average monthly active users across 130,000 user-governed communities, and each community is maintained by at least one and often many volunteer moderators.[4] Though Twitter, Instagram, and others use centralized, platform-driven moderation models supported by distributed labor of users

[2]Gengle's manual and Hiltz and Turoff's book provide strong evidence that the term "moderator" came to be used in online contexts because of the perceived similarity of the role to that of offline "conference moderators".

[3]https://web.archive.org/web/20191208191626/https://singjupost.com/full-transcript-mark-zuckerberg-at-facebooks-f8-2018-developer-conference/?singlepage=1

[4]https://www.redditinc.com/press, accessed 9 Feb 2020

who flag and report content, newer platforms such as Twitch and Discord rely heavily on volunteer, community-driven moderation.

Recent research has described processes of volunteer moderation in online communities, namely Reddit [4, 15, 29, 37], Twitch [37, 41], Discord [16, 19], and Facebook Groups [37]. Much of this work, e.g., [37], frames volunteer community moderation as a set of *processes* or *tasks* and emphasizes *actions*, like rule-making, content removal, and user removal, rather than the implicit values that guide overall approaches.[5] Building on examples set by Gengle and in other more recent work, I re-examine volunteer moderation from a metaphorical perspective, not as a set of duties but as complex set of *social roles*, asking questions like "How do volunteer moderators conceptualize what they do?", and "What values do they bring to their work?"

In this paper I present a map of the conceptual territory of volunteer community moderation roles across multiple platforms, using metaphors to illustrate how moderators make sense of their power to shape, guide, and influence their communities. I perform thematic analysis of 79 interviews with volunteer moderators from Facebook Groups, Reddit, and Twitch, including 56 first-round interviews and 23 follow-up interviews (conducted two years later) which focused on metaphorical language. I chose these three platforms because they are among the largest community-based sites and each relies heavily on volunteer moderators while, at the same time, featuring a significantly different culture and set of features. For example, Twitch is based around live video streaming with synchronous chatrooms and a gaming focused culture while Reddit is a primarily text based asynchronous platform, but posts can contain links to multimedia. Facebook Groups are a mix of text comments and multimedia posts and are often private or secret. I explored the metaphorical language that describes how moderators see their work and analyzed these examples to identify emergent categories. Each metaphor I present, from moderator as "Gardener" to moderator as "Police" to moderator as "Piñata", contains three core pieces of information: (1) The social role a

[5]Notable exceptions to this include [4], [35], [41], and [42]

moderator believes that they hold in the community; (2) An implicit set of values for what they believe their community should be like; and (3) An implied set of heuristics that shape how they make each moderation decision.

Understanding how volunteer moderators make sense of their roles paints a more complex and complete picture of their work. It also invites new lines of empirical research that illuminate relationships between platform cultures and moderators' values, and connections between moderators' value orientations and different community outcomes and characteristics (e.g., growth, conflict, supportiveness). These metaphors can also help platforms develop tools and spaces that support moderators far beyond the oversimplified but widely-presumed role of content removal and instead focus on their values, social preferences, and approaches to community support and management.

4.2 Prior work

I position this work at the intersection of literature on metaphors to understand diverse ways of thinking around abstract concepts, and the literature that discusses moderators' work practices.

4.2.1 Metaphors and social behaviors

Metaphors are essential to how humans understand the world. As Lakoff and Johnson write in *Metaphors We Live By*, "If we are right in suggesting that our conceptual system is largely metaphorical, then the way we think, what we experience, and what we do every day is very much a matter of metaphor" [21, p. 3]. They provide numerous examples of how metaphors can shape how we perceive phenomena that are part of our daily lives. For example, Lakoff and Johnson contrast two metaphors for acts of conversation: *war* and *journey* [21, pp. 78–82]. Phases of combat can be metaphors for stages of a confrontational conversation, e.g., planning a strategy, attacking, defending, reaching a stalemate, and agreeing to a truce.

Alternatively, a laborious conversation could be described with terms like setting out, going in the wrong direction, going in circles, and having come a long way. These two metaphorical framings are certainly not procedural descriptions of conversational acts; they instead illustrate different approaches to discussion.

Psychologists have studied the importance of metaphors in human cognition, showing connections between the metaphors we use to describe the world and internal cognitive representations. Psychologists have proposed *conceptual metaphor theory* to explain how metaphors act as a common cognitive tool to help people experience and make sense of abstractions, rendering difficult concepts more concrete and understandable [25]. According to this theory, metaphors both guide and are reinforced by individuals' thoughts, feelings, and behaviors. These metaphors influence social outcomes such as interpersonal judgments, problem solving strategies, and relationship satisfaction [26]. Metaphors can reveal the similarities and differences in individuals' mental representations of and perspectives on shared contexts, indicating their importance in social perception and coordination [23, 24].

Metaphors have a profound effect in guiding individuals' information processing and behaviors, framing one's roles and responsibilities as well as understanding personal and cultural values. [7] contrasts two dominant cultural metaphors in American and Japanese business contexts - football and the zen garden - to highlight how underlying values implicit in these metaphors (e.g., individualized specialization within teams in football; harmony and strict adherence to rules in gardening) illuminate how companies (and the individuals and teams within them) operate within each culture. [32] provides an analogous cross-cultural investigation of metaphors for nation-states and political bodies. Similarly, [30] revealed how the values underlying political ideology affected how liberal and conservative participants conceptualized the qualities of effective authority figures through metaphor. They found that liberals were more likely to use family metaphors of parental nurturance and fairness, while conservatives were more likely to use metaphors of discipline and rules enforcement in their depictions of authority. These studies show how metaphors and values are mutually

reinforcing: values more foundational in a given culture or ideology shape the metaphors used, and those metaphors can be a powerful, validating expression of those values.

The metaphors I present in my analysis capture different outlooks and techniques to moderation. As Lockton et al. write, "metaphors are not the thing itself–they are always an abstraction, a model of the situation [...] They can be a map to a territory, but should not be mistaken for the territory." [27, p. 322]. I believe that a thorough map to this territory is valuable for more than just understanding moderation's central role in mediating online communication. The metaphors I present can help explain central tenets of moderator actions, such as how norms and practices are developed, how on-boarding of new members will occur, and how tensions and disagreements are overseen.

4.2.2 Moderation in online spaces

Research on moderation has ranged from a functionalist definition of moderation as the processes surrounding removal of content and bad actors to a perspective emphasizing the evolving intersection of norms, attitudes, and context for moderation. Our work takes the latter approach to complement recent more functionalist studies of community self-moderation.

Beginning in the 1980s and early 1990s, early research on moderation in online communities used metaphor-heavy framings and lenses to observe contextual factors (e.g., [33] and [39]). This work was frequently qualitative, often theory-driven, and primarily ethnographic. It focused on online communities in spaces like Usenet, Multi-User Dungeons (MUDs), and in some cases Internet-Relay Chat (IRC) and mailing lists, which were primarily governed by users. While researchers did at times analyze moderation from a procedural, content removal-based lens, these analyses were typically situated within a narrative describing a larger theme or context of operation, e.g., "Hierarchies of power" [33, pp. 118–120] or "Diversity as a source of conflict" [39, pp. 143–146].

Later work shifted toward functional analyses of moderation practices and processes. Despite the growth of corporate social media like Myspace and Facebook, this research

still modeled approaches to moderation and regulation of online behavior in user-governed communities like Slashdot, Wikipedia, the MovieLens platform, and free and open-source software communities. This work primarily used quantitative methods and focused on metrics for success and growth. For example, [22] used quantitative analysis of users' moderation logs to show the strengths and weaknesses of Slashdot's distributed moderation.

Recent research has shifted to focus on moderation from the platform's perspective. Legal scholarship in particular has taken a functionalist, removal-based approach, though recent work has expanded to emphasize the importance of alternative approaches to handling problematic content that extend beyond removal [5]. In her "The New Governors: The People, Rules, and Processes Governing Online Speech", [20] documented the "governance" processes of social platforms, describing how processes have evolved and noting the influence of certain ingrained philosophies. [6] and [2] focused specifically on the moderation of nonconsensual pornography, discussing the current state of policies and laws surrounding whether and when it should and could be removed from social platforms. Research on community self-moderation has continued to work from a similar functional perspective – for example, [37] presented three processes for volunteer moderation containing 15 steps and a total of 45 themes and variants. While they include some contextual pieces in their model, including "Development of a moderation philosophy", their work describes moderation as a series of possible actions.

Though these works address users' experiences and values, each emphasizes the *actions* and *processes* in moderation. Other work has focused on the broader social context from a more phenomenological perspective. In *Custodians of the Internet*, Gillespie documents actions for moderation, including removal, filtering, suspension, recommending, and curating [12, pp. 207–208], but he situates his descriptions within narratives of the pressures platforms face. Norms and values engage with power dynamics and legal restrictions to create a progression of rhetorical, technical, and organizational approaches to being *custodians*, the core metaphor of Gillespie's work. Similarly, Matias used a contextual lens to explore the

labor of Reddit moderators using the term "boundary work", which he drew from [10] to describe the evolution of moderation from free labor to civic upheaval then towards oligarchical moderation practices [29, pp. 2–4]. [40] used a similar perspective to analyze users' experiences, examining the impact on users of opaque moderation actions taken by platforms like Facebook, and [9] also looked at users' behaviors in the context of systems of moderation on Instagram, identifying ways in which they attempted to circumvent removal processes.

More recently, [18] explored Reddit moderators' perspectives on transparency in moderation processes including rule-writing, removal notifications and explanations, handling appeals, and general rule-enforcement. Their work refined Matias's concept of moderator oligarchy, identifying the reasons moderators use or reject transparent procedures. In an explicitly metaphorical analysis, [41] categorized Twitch moderators' roles into "Helping Hand", "Justice Enforcer", "Surveillance Unit", and "Conversationalist", each accompanied by implicit goals like maintaining civil discussion and fostering supportive interactions. [35] focused primarily on commercial content moderators who work for platforms specializing in conversation forums, but identified significant overlap between these moderators and volunteer community moderators in ways that relate to a "logic of care". These authors see moderation as a complex and evolving interplay between many stakeholders.

I build on this prior work to provide a more comprehensive categorization of the social roles in volunteer moderation. I present a detailed set of role-categories sensitive to the complex contexts of volunteer moderation across different platforms. As platforms have evolved, both in who participates in moderation and what activities and roles are included, a reexamination of cross-platform volunteer moderation efforts through the use of metaphor is valuable to map the current state of moderation and to ground discussion about possible future states.

4.3 Methods

I performed 79 semi-structured interviews of volunteer moderators, including first-round interviews with 56 moderators from Fall 2016 through Spring 2018, and follow-up interviews approximately two years later with 23 of these interviewees. The moderators I interviewed were from Facebook Groups (15 first-round, 7 follow-ups), Reddit, (21 first-round, 9 follow-ups), and Twitch (20 first-round, 7 follow-ups).[6]

Interviews typically lasted between 30-60 minutes, with variance based on the number of communities moderated, depth of engagement in these communities, and interest in meta-discussions about moderation. Interviewees were compensated 15 USD (or the equivalent amount in their local currency) for participating. Interviews were performed in English, and while the majority of the interviewees were from the United States, eight out of fifty-six were from the UK, two were from each of Canada and France, and one each was from Sweden, Mexico, Australia, and Germany. I intentionally oversampled from underrepresented populations when possible in order to amplify the voices of these populations. I also hypothesized that a more diverse sample might lead to more variety in the metaphors uncovered. For example, while, as of mid-2019, Twitch's advertising website reported that 81.5% of its user base was male,[7] our sample of twenty interviewees from Twitch contained five cisgender and two transgender women.

Initial interview recruitment combined direct messaging to moderators and snowball sampling to recruit a diverse set of group sizes and themes represented. Follow-up interviewees were recruited through re-messaging the moderators via the same channels they were originally contacted. Though I believe a response rate of 40% (23/56) is reasonable after a two-year delay, I recognize that there is likely bias in which interviewees responded to re-interview requests. As such, I do not make claims that these results are statistically representative of

[6]Prior work reported on the results of the first-round interviews from a procedural, functionalist perspective, which did not focus on metaphors [37].

[7]https://web.archive.org/web/20190629052039/https://twitchadvertising.tv/audience/

all moderators on these platforms.

The goal of performing follow-up interviews was twofold – first, to explore specific topics such as metaphors and philosophies for moderation, but also to capture how moderators felt their roles had evolved as platforms had added new features and changed their policies. In analyzing our first-round data, metaphors emerged as an implicit theme, with 51 of the 56 interviewees using metaphorical language at least once to describe their work or situations they found themselves in. While some moderators used metaphors to describe problematic users, e.g., calling them "garbage" that needs to be taken out (a metaphor also used by [8] nearly forty years prior), most metaphorical language described moderation styles and roles. I designed follow-up interview questions to investigate moderators' philosophies and their origins. Metaphorical language was not used evenly across interviewees from the three platforms; Reddit moderators used many more metaphors than moderators from Twitch or Facebook Groups, even without prompting. Twitch moderators used less metaphorical language both when unprompted and when explicitly offered the chance to do so. I explore possible explanations for this in our Discussion.

After completion and transcription of the follow-up interviews, I performed thematic analysis following Braun and Clarke's guidelines and process [1]. Our units of analysis were passages with metaphorical language where moderators described how they viewed their roles. I identified 244 chunks of text that used metaphorical language in our first round interviews and 53 chunks in our follow-up interviews. Having familiarized myself with the data, I performed open coding, abstracting from, e.g., "I had been moderating for her on YouTube where that was a constant battle" (T15) into "Moderation as combat". Next, two researchers independently grouped these abstractions into broader sub-themes, including moderators as legislators, mediators, combatants, and police. The researchers then met and performed affinity diagramming to consolidate and group these variants, arriving at five social roles with three to six variants in each group. I then checked these themes against the whole dataset and solicited feedback from several reinterviewees about our findings.

4.4 Results

Our analysis surfaced five social role categories with 22 metaphor-based variants, shown in Figure 4-1. Note that these roles are not *functional* roles; while [37] found little formal division of tasks between moderators beyond hierarchy and technical vs. non-technical work, I found significant differences in the *social roles* that moderators felt they played in their communities as expressed through metaphor. These metaphors indicate diverse value sets, which moderators articulated through conversations about how each metaphor reflected their relationships toward their communities. These metaphors also contain implied heuristics that shape action and decision-making. In this section, I describe the interplay between these characteristics and how they connect to social roles.

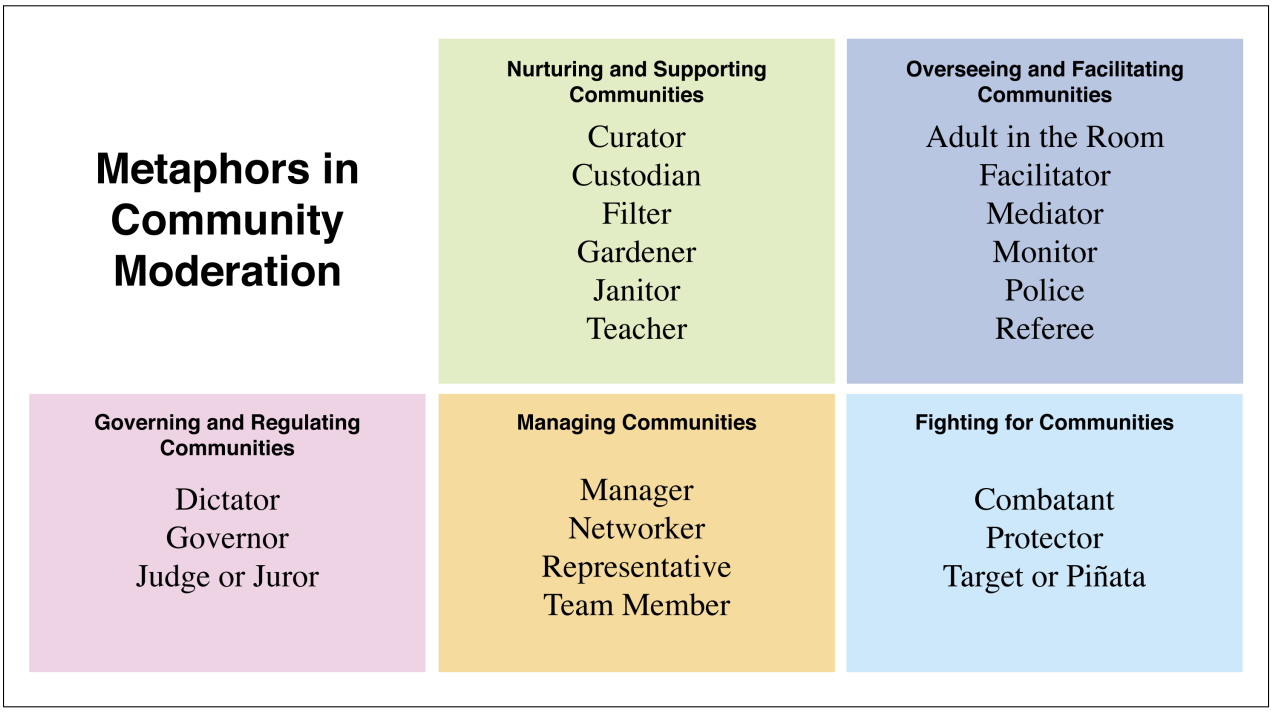


Figure 4-1: Five categories and twenty-two variants of metaphors for roles in volunteer community moderation. The orientation of categories and metaphors is not indicative of any hierarchy or relationship between groups.

4.4.1 Nurturing and Supporting Communities

> "The moderators that I get along with most at an archetypal level tend to the space, meaning they tend to the people, they tend to the vibe... they are socially aware of what emotions are moving through the space at any given time and they are tending those emotions and the actions tied to them." – T2, reinterview

> "My way of looking at it is that in a way you're kind of leading the group like a teacher would, like a class. You're making sure that everyone feels safe, you're trying to keep a safe environment for everyone and trying to keep them learning and updated on new things that are coming into the group." – F4, reinterview

Moderators who fit the *Nurturing and Supporting Communities* category take care of and nurture a community, with varying emphasis on care versus maintenance and active versus passive engagement. First, the **Gardener** nurtures and "tends" to the community, both in "planting seeds" of conversations and interactions and, as F7 described in their reinterview, "pull(ing) weeds" – problematic content and disruptive users – in order to make space for more positive interactions, similar to the nurturing attitudes of moderators highlighted by [42] and the "Helping Hand" Twitch moderator archetype identified by [41]. Gardeners see themselves as slightly detached from individual social interactions in a community, focusing instead on larger community goals. **Custodians** show similar feelings of responsibility and a desire to help a community grow, likening themselves to a detached "steward of their community" (R11). Several interviewees explicitly contrasted Custodians with **Janitors**, where the Janitor role was associated with language like "trash", "waste", and "repair[ing] broken windows and sweep[ing] the floor" (R13). The most passive type of moderator in this group, the **Filter** moderator behaves similarly to a physical filter, taking away what one moderator called "low effort content" (F13) without intervention. Filter moderators are much less likely to engage directly with community members; instead, they put up "pre-screens" (F6) to identify problematic users or content before they enter a community.

A **Curator** works with a more specific vision and a sense of expertise in the community's interest areas compared to Custodians or Janitors. Curators articulate their roles as more "editorial", like "a newspaper editor" (R8, reinterview), who actively guides both the content in the community and its members' behaviors rather than guiding growth. Moderators who used Curator-related language were often associated with a knowledge-based group, where they may have been chosen as moderators for their expertise. The final metaphor, **Teacher**, is also associated with expertise but implies a more active, hands-on approach to moderation. Teachers see their community as a "classroom" (F4), with at least some community members being "students" who can be "trained" (F7) to behave in particular ways. One moderator directly compared moderation work to their experiences as a school teacher.

Moderators whose roles are described in the *Nurturing and Supporting Communities* category envision a community that is clean, trained, and often evolving over time. Correspondingly, these metaphors describe heuristics for moderation decisions based on their vision for the community. There is significant diversity within this category. The distinction between Gardeners, Teachers, Curators, and sometimes Custodians, who have a vision for what they would like their community to be, and Filter and Janitor moderators, who have a vision of what they would like it *not* to be is comparable to the stark differences identified by [35] between commercial content moderators who are forced to operate like machines, filtering through content and removing it when necessary, and moderators who have the resources to nurture communities. It is important to note that, while the Filter and Janitor moderators in our dataset chose these approaches to moderation, the commercial content moderators studied by Ruckenstein and Turunen were forced into reactive, filtering roles because of the way the platforms structured employees' work processes.

4.4.2 Overseeing and Facilitating Communities

> "A lot of times mods will just kinda sit back and will see how far it progresses, kind of like bystanders, to see if this arguing match becomes a physical fight, and

then they have to step in." – T2, first interview

"If the ethos of the group itself is more concerned with like, more democratic engagement, I usually try to be a diplomat. And when I see conversations getting out of hand, I'll try and help the people who are arguing understand each other better." – F11, reinterview

In contrast to the above six metaphors, which focus on long-term vision, the metaphors in the *Overseeing and Facilitating Communities* category view moderation as interpersonal management and intervention. **Mediator** and **Referee** moderators hope to resolve disputes. Mediators often step in when discussions get "heated" to help "cool things down" (R2). Mediator moderators help community members reach equitable resolutions to their disagreements, ideally with minimal animosity to "help [users] understand each other better" (F8). In contrast, Referee moderators resolve disputes by referencing rules or a body of accepted knowledge, with the goal of being "fair" and "neutral" (F7) in application of these rules; one interviewee even used the sporting metaphor of "putting a yellow flag" (T2) on a person for violating rules. When a Referee steps in, their goal is to decide who is right and wrong in a given situation and to support their decision with clear justification. While a Mediator's legitimacy comes from their ability to de-escalate, a Referee's legitimacy comes from making fair and consistent decisions. Both of these metaphors match well with the conception of a moderator in [13], which comes from traditional conference moderators in academic conferences.

Similarly, moderators who see themselves as **Police** and the **Adult in the Room** are expected to make relatively consistent decisions but articulate their reasoning less frequently and in less formal ways. Police were described as "guard dogs" (T2) comfortable making in-the-moment decisions about rule-abiding behavior and taking corresponding (and often quick) action. This matches many of the features of the "Justice Enforcer" moderator persona identified by [41]. One moderator said that, in response to offensive behavior, "of course we kick them out instantly and we don't tell them anything" (F2). Adult in the Room

moderators view misbehavior as resulting from immaturity or childishness that they must contain, sometimes via lecturing users on proper behavior. These moderators try to maintain order amid chaos with language around physical space, with one moderator describing unruly Twitch communities as a "crazy house party". In their reinterview, F11 described the mix of patience and directness required to be an Adult in the Room – "I built this house, I invited you in, and you track mud all over my carpet [...] Nah, get the fuck out or clean up your mess."

The **Facilitator** and **Monitor** archetypes are more passive variants of the above roles. Monitors are quiet observers, someone who "would watch and would report if something bad happen[s]" (R6), akin to Wohn's "Surveillance Unit" [41]. These moderators note potential issues and mark users for review, stepping in only if absolutely necessary. Facilitators described themselves as "relaxed" about moderation, frequently referencing the effectiveness of community self-regulation. Several Reddit moderator interviewees described allowing users to self-moderate, "letting the upvotes and downvotes do their thing" (R5). In some cases, Facilitators act as a "host" to encourage socialization and conversation in their communities.

In contrast to the *Nurturing and Supporting Communities* category, moderators in the *Overseeing and Facilitating Communities* category expressed philosophies for ideal social interaction, both between themselves and communities and inter-community interactions. For example, several interviewees who moderated spaces for political discussions on Facebook known as "discourse groups" wrote extensive rules and expectations for appropriate modes of interaction. Correspondingly, these moderators' heuristics for intervention evaluated whether a piece of content or a conversation would be disruptive to the community, a philosophy more akin to "firefighting" than gardening.

4.4.3 Fighting for Communities

> "There's periods of contention and intense discourse where I'm either the most loved or hated person. Which sounds so egotistical, but it's like being a piñata

sometimes." – R4, first interview

> "In everything that I do, I try to carve a safe space for queer people, for trans people, especially for young trans kids. I want to do everything I can with my life to make things better for them, for a safe, inclusive and diverse space, and I am willing to take any measure necessary to maintain that. If you are making things unsafe, harmful, mean, negative... you're gone." – T1, reinterview

For moderators who describe their social roles as *Fighting for Communities*, moderation is a battle. The first type of moderator in this group, the **Combatant**, sees active conflict as part of their job. These interviewees used metaphors like "sniping" problematic content (T20), having an "itchy trigger finger" (R13), and being "drafted" (R4) to manage conflicts in communities. This militaristic language was most frequently used when discussing content removal and bans; moderators rarely mentioned warning users, giving explanations for removal, or socially engaging with their communities. Moderators using the **Target or Piñata** metaphors described a similar philosophy but focused more on the pushback they would receive from angry community members; one moderator directly described being beaten up by their community as being like a "piñata" (R4). Another moderator described "hav(ing) to put your flame suit on" after difficult decisions. These moderators do not feel obligated to have a discussion with or accept feedback from those who disagree; rather, they stand by their decisions and try to weather the proverbial storm. Note that this metaphor was not used by moderators who were targets of harassment themselves because of, for example, their race or gender; moderators who described themselves as a Target or Piñata chose this role willingly because they saw it as their job to absorb complaints as part of the moderation process.

The third metaphor, the **Protector**, describes a moderator who is invested in creating a safe space for community members. These moderators often seek to provide a space for underrepresented groups or vulnerable users, e.g., "young trans kids" (T1). Unlike the Target or Piñata metaphor, these moderators had typically experienced harassment personally, which

led them to want to protect others like them from having the same experiences. For this reason, Protectors are quick to remove anyone who poses any perceived risk to the safety of their community (F6), akin to the "martyr" identified in [29], who adopted the role of a defender of her community (p. 5).

All three metaphors highlight the processes surrounding content removal. None address, for example, education of offenders, reconciliation, or the idea of fairness or transparency. Moderators who are *Fighting for Communities* unapologetically follow their values in making decisions about what and who to remove.

4.4.4 Managing Communities

> "So what I wanted to do was hire people with different opinions that are good at different positive things. We hire [new] moderators quite often. We don't demote [the old ones] because they're terrible, but because they're, you know, they got busy with life and whatnot." – R3, reinterview

> "We just put a little job application, if you will, on the pinned post." – F12, first interview

Much has been written about content moderation as labor, both within commercial and organizational [12, 34] and social and civic structures [4, 29, 41]. Many interviewees, with a strong majority from Reddit, described approaches to moderation with organizational and employment metaphors. The most common of these was the **Team Member**, where moderators described an organized group of "coworkers" who were "recruited" (R8), "hired", and occasionally "promoted" or "demoted" (R3). These metaphors were used despite no formal employment status or financial compensation for their work, aligning with the description of Reddit moderator labor in work by [29] and contrasting with the formal employment structures of the commercial content moderators described by [35], some of whom had previously been volunteer community moderators. These moderators view their work as a second job,

with regular hours spent on tasks, like working through moderator private mail on Reddit. These hours often overlap with their "first" job's hours, with moderators taking advantage of downtime at work to interact with their community. Some teams have a **Manager**, a user who is often recognized as a "head mod" of a given community, who organizes the efforts and "creates infrastructure" for others. In our interviews, the Manager metaphor did not have political governance undertones; instead, this role was articulated in terms of necessary labor and management of organizational processes.

Other metaphors more directly explored the relationships between moderators and their communities and other spaces. A **Representative** moderator is the "main face" (F12) of a moderation team, whose job is to provide a positive impression of the moderators to the community akin to "public relations". Several Twitch moderator interviewees described themselves as "representatives" of the streamer they moderated for. The fourth and final metaphor, the **Networker**, describes moderators whose goal is to make connections with different communities for both personal and "professional" reasons. These moderators act as "liasons" to share tools and strategies and to coordinate actions. For example, one moderator recalled a large discussion between sports-related groups about handling controversies involving the American football player and activist Colin Kaepernick. However, several Reddit moderators described a negative version of this Networker role – "cabals", where users accumulate personal power by becoming moderators on many large subreddits, as previously discussed by [28].

These organizational metaphors highlight similarities between user-driven moderation and platform-driven moderation in approaches to work and labor. They also reveal an underlying tension between the commercial and social natures of self-governed communities. Despite the fact that many Twitch communities are a significant source of income for the streamers, no interviewees from Twitch described the streamer as a manager, and very few used organizational metaphors other than Representative. This may reflect these moderators' desire to view their relationships in friendly, social, and supportive terms, but this approach

can lead to conflict when commercial and personal relationships collide [41].

4.4.5 Governing and Regulating Communities

> "We tried public referenda. Being a political subreddit, people get really ugly about these rule changes so we try to be really democratic about it if we can. Sometimes it works, sometimes it doesn't. Sometimes I just issue edicts and that's that." – R4, first interview

> "When I logged into the back room of uh, the [subreddit] today, I saw that they were actually trying a case basically of whether or not to ban some guy. So there it looks not like an individual judge but more like the supreme court, Right. So you got a panel of judges." – R8, reinterview

The final category of metaphors and social roles draws from legal and political governance. In contrast with the *Managing Communities* or the *Overseeing and Facilitating Communities* metaphors, these roles describe moderation as a form of government, interpreting between the "letter" and "spirit of the law" (R2) and sometimes "providing amnesty". The **Judge or Juror**, for example, is a moderator who decides on a case-by-case basis whether an action warrants punishment, either individually as a Judge or as part of a Jury of moderators. Moderators who saw their roles this way framed rules around concepts like "common law" and "civil law" and described trying to go "by the book" in adjudicating decisions (R8). This metaphor contrasts with the previously-discussed Referee, where the Judge or Juror examines the behavior of a single user and the Referee attempts to determine whether one party or another is correct.

While the above represents a judicial branch that manages disputes, moderators also discussed variants of an executive branch. A **Governor** leads with a general sense of consent from the "governed", though typically not through a democratic mandate. Governors, sometimes described in terms like "president" (R2), usually discuss decisions with their mod-

erators and take other opinions into account, but have the final say in matters of moderation where they may "line-item veto" new policies. Moderators also used the term **Dictator**, but this was applied only in rare cases when a moderator exerted a sort of "tyrannical" authority over a community that did not accept their legitimacy. In a few cases, being a Dictator described the need for one person to make a final decision, an executive "fiat", when issues became too contentious for agreement to occur. In other cases, moderators pointed out that communities unfairly accused them of being "Nazi mods", with the perception that they had over-enforced rules. This comment was most common among Twitch moderators, who described the widespread use of the term "Nazi mod" on the platform.

The metaphors used by moderators here are inspired by real-world governance structures. While Dictators represented a more autocratic form of governance, Judge or Juror moderators formed a judicial branch of a more democratic government, complemented by an executive branch in the form of Governor moderators. However, no metaphors were used that classified moderators as members of a legislative branch or even as elected representatives of the community. Though these moderators may have governed with democratic ideals in mind, their governing structures were more akin to an oligarchy than a democracy. [38] described similar oligarchic concepts on Wikipedia, and [29] noted the presence both of oligarchs and self-described "dictators" on Reddit (p. 5).

4.4.6 Establishing Face Validity: Feedback from Interviewees

After constructing this taxonomy, I sent it to interviewees for feedback to test its face validity. Most expressed approval, e.g., "That is actually very accurate. I've been in groups that are run like all of these or a mixture of 2." (F4). Several moderators suggested new names for our variants or additions to the framework. For instance, R3 suggested using "Dictator" instead of "Tyrant" (which I had originally used), and I made this change in the final framework. R11 asked whether "Steward", which they had mentioned in their re-interview, would fit in the framework; I elected not to formally incorporate this as no other interviewees had

mentioned it, but I added it as an example in discussing the Custodian metaphor. Finally, R8 suggested a additional category for **Media**, **Journalist** and **News-editor** metaphors. I did not adopt these metaphors directly as they were not present in our dataset, and some overlapped significantly with the social role of Curators (with one moderator explicitly comparing a Curator to a "Newspaper-editor"), but I believe these metaphors may merit additional consideration in future work.

4.5 Threads for future research

The twenty-two metaphors I present here paint a nuanced picture of the social complexity and roles of volunteer moderation in online communities. While recent work has focused on moderation as a process built around rules and removal – i.e., what to remove, what not to remove, and how to communicate this – removal is only one piece of a deeper social process of nurturing, overseeing, intervening, fighting, managing, governing, enduring, and stewarding communities. Our interviewees described these roles as fluid and changing; the median interviewee self-described using three different metaphors spanning two of the five categories. Though our sample is not sufficient for rigorous quantitative evaluation of overlap of different metaphors, prior work offers hypotheses for how clusters of metaphors may be interlinked. For example, situational factors are likely involved in these determinations; as suggested by [14], metaphors can drive action, and the current social context serves as a powerful constraint to activate particular metaphors that best suit one's goals or motivations. Moderators might shift their metaphorical lens depending on the nature of the transgression or the characteristics of the transgressor (and, indeed, the transgressor's own metaphors for governance and authority).

Beyond situational factors, the characteristics of moderators – their demographic and social identity characteristics, their personalities, and their belief systems – undoubtedly shape their views on authority, community, and governance. As noted previously, political

ideology shapes the types of parental metaphors invoked about political leadership [30]. Prior work has also revealed how gender and one's endorsement of and adherence to gender roles can affect one's choice of agency- versus nurturance-oriented metaphors in developing self-concept, and how personality traits can influence the metaphors used in thinking about oneself and others [31]. Various facets of the self-concept may be similarly impactful in influencing what metaphors are used to think about the roles and duties of moderation.

Certain metaphors were far more common among moderators from specific platforms, suggesting that platforms' affordances, histories, and social norms may also impact the metaphors moderators adopt. For example, *Nurturing and Supporting Communities* metaphors were used by Reddit and Facebook Groups moderators much more than Twitch moderators, while *Overseeing and Facilitating Communities* metaphors were more frequent on Twitch. Twitch uses an active, synchronous live chatroom, so it may be more difficult to "curate" a moving conversation or "plant" ideas, while the slower, asynchronous communication on Facebook Groups and Reddit may allow moderators more time to consider longer-term plans and goals. Perhaps because of the politically-charged aspects of Reddit's culture and history [28], *Governing and Regulating Communities* metaphors such as Governor and Judge or Juror were used much more by Reddit moderators. Similarly, the greater use of the "Adult in the Room" metaphor by Twitch moderators may reflect the perception that many Twitch users are young and that their problematic behaviors are a result of immaturity. The low volume of metaphorical language from Twitch moderator interviews may result from the youth of Twitch as a platform; as [3] point out, metaphor use can increase as expertise in a field grows, so it is possible that conceptual models for moderation on Twitch are still maturing.

Our analysis catalogues the social roles found in our dataset, connects them to metaphors in prior work, and makes more explicit their usefulness in understanding moderation practices and relationships. Consistent with prior work [21, 23, 24], metaphors provide explanatory power in grounding behaviors, choices, and values of moderators in decision-making and

everyday interactions. For example, the difference between Governors and Curators implies differences in perceived relationships, behaviors, and values that moderators hold. Governors may look towards the rule of law in legitimizing decision-making, seeking approval from other moderators and the community, while Curators possess a vision of community vibrancy, health, and success and foster this vision in their interactions.

These metaphors also illuminate threads for future work. For example, a cross-platform "census" survey of metaphors would offer validation of this framework and could explore how cultural differences impact the use of metaphors for moderation. More in-depth exploration of how each category and perhaps each metaphor manifests in practice could also deepen understanding of the space. Understanding the complex interplay between individual, social, and contextual variables in shaping metaphor choice and deployment is a crucial next step in this line of inquiry; I also acknowledge that these same variables affect other cognitive processes beyond the activation of metaphorical frames. Future work taking a more holistic view of moderator cognition should consider the role metaphors play relative to other psychological mediators (e.g., stable or temporarily activated attitudes, norms, rules, emotions, etc.) in shaping moderators' perspective-taking and decision-making processes.

4.6 Threads for Design

Though much of this chapter has focused on the creation of the framework as a contribution to an ongoing academic discourse, metaphors can also be generative in helping designers and developers introduce new concepts. As in Lakoff and Johnson's examples of metaphors for conversation, the application of a new metaphor to a situation where it does not immediately make sense can facilitate new perspectives [21, pp. 78–82]. For example, in Lockton et al.'s *New Metaphors* method, designers are tasked with combining names of phenomena and abstract concepts with photos of real-world objects. They then extract metaphors from these often unorthodox combinations that can then be used to inspire ideas for new products,

services, and interfaces, and also to help reframe complex situations [27, p.323]. Existing systems for moderation can be connected with some of the metaphors I present; most existing tools for moderation align with a small subset of the metaphors described here. For example, Twitch's timeout and ban features align well with Adult in the Room or Police metaphors, while Reddit's AutoModerator and Facebook Groups' screening questions tie closely with Filter metaphors. Taking into consideration less commonly used-metaphors could help give designers a new perspective on what moderation could look like in the future.[8] Here I present two examples of how metaphors could be translated directly into new types of tools to assist community moderators.

Example 1: MetaQueue

Whether on Twitch, YouTube's livestreaming, Facebook Live, or even on adult livestreaming platforms like Chaturbate,[9] a common challenge in larger streams is bringing the streamer's attention to the messages sent by users that need or deserve their attention. On Twitch, messages sent by users may contain reasonable questions or good suggestions, but could be lost in the flow of messages because the streamer cannot read every message and still focus on playing their game. Similarly, popular webcam performers may miss legitimate questions or even monetary tips that they need to acknowledge or respond to because of the chaotic and fast-moving nature of their chatroom.

In order to address this, I propose *MetaQueue*, a real time *Filter* metaphor inspired tool that allows streamers' moderators to directly select which messages will be entered into a separate queue on which the streamer will place their primary focus; the streamer may occasionally look at the general chatroom, but their attention will be primarily focused on the special curated MetaQueue, which contains messages that their moderators deem most important for them to respond to.[10] Users would not know that their message was selected

[8]The BabyBot work I have done separately from this dissertation looks at creating an environment where a moderator can serve as a Teacher or an Adult in the Room [36]

[9]See Jones's work on adult livestreaming for details about moderation challenges on those platforms [17].

[10]Note that, while I frame this as a task to be done by human moderators, it could also be done perhaps passably well by well-trained algorithms.

to appear in the MetaQueue, and the MetaQueue's existence would not necessarily even be visible to the audience as a whole; they would only be aware of the fact that certain messages were getting a response and others were not.

On its surface, this does not look particularly like a moderation tool. It seems more like a filter designed to maximize efficiency of streamer-chat interaction by highlighting the most important messages. However, this in itself serves a valuable moderation function – the choices a streamer (and/or her moderators) makes in determining which messages to respond to set a very visible standard for what types of behavior are most welcome in a space. For example in their reinterview, T2 noted that one of her primary moderation strategies involved "putting someone on ice". This meant that, when a user had begun to behave in an unwanted way, they were simply ignored and other users would have their comments responded to and questions answered. T2 reported that users very quickly learned what types of engagement were most likely to get positive attention from her. MetaQueue would serve a similar function. In addition to making sure streamers saw messages they needed to see, it would also make sure streamers were responding to (and only to) messages that followed desired behavioral patterns. Other users' messages would either be dealt with by moderators or simply ignored. This sort of "pseudo-passive" moderation could be a quiet but effective way to specify norms for behavior in a livestreaming space.

Example 2: The Target Dummy

In many of the 79 interviews I performed for this project, moderators brought up the experience of being flamed or verbally abused by users who had been punished in their community. Moderators often took this for granted as a part of their job, but many noted that this abuse could take a toll on their emotional state in both the short and long term. These moderators felt like a Target, or a Piñata. In order to develop a tool that would address this situation, the emotional load on moderators would need to be decreased while still providing any necessary information to offenders (e.g., information about what rule they had broken, an explanation of their punishment, or a procedure for how they could be

readmitted to the space if applicable).

In order to address these dual challenges, *The Target Dummy*, a chatbot designed to handle the abuse that moderators would otherwise receive. When in a hostile situation, a moderator would start an instance of The Target Dummy to which the offender would immediately be referred (whether by starting a new conversation window or simply launching the chatbot within the existing conversation instance, temporarily taking over the moderator's account). The Target Dummy would be a rules-based chatbot, akin to a customer service hotline that acts as a stand-in for the moderator, that would be launched with some basic information about the situation through a command-line style prompt, e.g., "/run TTD offendername rulesbroken punishmentlength pathtoredemption", or "/run TTD jseering 3,4 24, 1". This might indicate that user jseering has broken rules 3 and 4, has been banned for 24 hours, but can be readmitted through process 1, which could be, e.g., some form of admitting mistakes and apologizing. The bot would be initialized with this information, which would then determine its responses and potential conversation paths. It would still be possible to curse or yell at the bot, but doing so would accomplish nothing and would not take any emotional toll on a human moderator.

One could argue that a user with questions should automatically be referred to some sort of customer chatbot, with access to a moderator as a last resort when their question cannot be answered by the bot, but placing the bot as a backup for problematic cases allows moderators to maintain the human element of interaction with as many community members as possible. Thus, The Target Dummy would be a quick way to offload emotionally-heated exchanges while still allowing moderators to engage with community members who are acting in good faith.

These two examples are brief exploratory concepts that show some of the breadth of the potential future space for metaphor-inspired moderation tools. There are many more avenues to explore both these two metaphors and the others through research and design.

4.7 Conclusions

Although our focus is on social roles as seen by volunteer moderators, there is potential for future work exploring the ecosystem of platforms and community members that support moderation. Platforms may benefit from reflecting on the metaphors implicitly embodied in the way they engage human moderators as well as explicitly embodied in automated moderation schema and policy documentation. Does Reddit see its volunteer moderators as Team Members or Janitors? How do platforms conceptualize the non-moderator community members who also facilitate moderator labor through flagging and reporting mechanisms? Are automated moderation algorithms supposed to act like Filters or Police, or should their metaphorical roles complement the work of human moderators in a more sophisticated way? Though the moderators discussed in this chapter have all been volunteer community moderators, this latter question connects directly with how the roles of commercial content moderators are conceptualized. [35] argue that commercial content moderators are "trapped in a cycle of responding to one post at a time rather than offering a meta-perspective to the discussion by overseeing and nurturing it" (pp. 1039–1040) in part because of the ways in which proponents of AI-based moderation systems have framed these algorithms as replacements for humans rather than assistants. Metaphors like the ones described in this chapter could support conversations about ways in which algorithms could complement the unique skills of human moderators.

Metaphors must be used judiciously. Their flexibility and open-ended nature can obscure important realities. In "The politics of 'platforms"', Gillespie argued that the use of the 'platform' metaphor in describing sites like YouTube makes it difficult to attend to the consequences of these companies' interventions into public discourse [11, pp. 359–360]. It is tempting to forcibly apply metaphors and categories to situations where their use may not be warranted, and I caution against using these metaphors to overgeneralize or to design systems that take agency from users and moderators. Instead, I hope that these metaphors

will illuminate the breadth of practices in volunteer moderation and allow us to see them from new perspectives. Building from the flexibility of these metaphors, we can better understand the current state of online moderation and think more openly about its future.

Bibliography

[1] Virginia Braun and Victoria Clarke. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, pages 57–71. American Psychological Association, Washington, DC, US, 2012.

[2] Danielle Keats Citron. *Hate Crimes in Cyberspace*. Harvard University Press, Cambridge, MA, USA, 2014.

[3] Nancy J Cooke and Michael C Bartha. An Empirical Investigation of Psychological Metaphor. *Metaphor and Symbolic Activity*, 7(3-4):215–235, 1992.

[4] Bryan Dosono and Bryan Semaan. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[5] Evelyn Douek. Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility. *N.C. J.L. & Tech*, 21:1–78, 2019.

[6] Mary Anne Franks. "Revenge Porn" Reform: A View from the Front Lines. *Fla. L. Rev.*, 69:1251–1337, 2017.

[7] Martin J Gannon. Cultural metaphors: Their use in management practice and as a method for understanding cultures. *Online Readings in Psychology and Culture*, 2002.

[8] Dean Gengle. *Communitree*. The CommuniTree Group, San Francisco, CA, USA, first edition, 1981.

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

[10] Thomas F. Gieryn. *Cultural Boundaries of Science: Credibility on the Line*. University of Chicago Press, Chicao, IL, USA, 1999.

[11] Tarleton Gillespie. The politics of 'platforms'. *New Media & Society*, 12(3):347–364, 2010.

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

[13] Starr Roxanne Hiltz and Murray Turoff. *The Network Nation: Human Communication via Computer*. MIT Press, Cambridge, MA, USA, 1993.

[14] Hans IJzerman and Sander L Koole. From perceptual rags to metaphoric riches—Bodily, social, and cultural constraints on sociocognitive metaphors: Comment on Landau, Meier, and Keefer (2010). 2011.

[15] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5):31:1–31:35, July 2019.

[16] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):55:1–55:23, November 2019.

[17] Angela Jones. *Camming: Money, Power, and Pleasure in the Sex Work Industry*. NYU Press, New York, NY, USA, 2020.

[18] Prerna Juneja, Deepika Ramasubramanian, and Tanushree Mitra. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. In *Proceedings of the 21st International Conference on Supporting Group Work*, New York, NY, USA, 2020. ACM.

[19] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. Technological frames and user innovation: Exploring technological change in community moderation teams. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):44:1–44:23, November 2019.

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

[21] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 2008.

[22] Cliff Lampe and Paul Resnick. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 543–550, New York, NY, USA, 2004. ACM.

[23] Mark J Landau. *Conceptual metaphor in social psychology: The poetics of everyday life*. Routledge, 2016.

[24] Mark J Landau. Using metaphor to find meaning in life. *Review of General Psychology*, 22(1):62–72, 2018.

[25] Mark J Landau, Brian P Meier, and Lucas A Keefer. A metaphor-enriched social cognition. *Psychological bulletin*, 136(6):1045, 2010.

[26] Annette Lareau. *Unequal Childhoods: Class, Race, and Family Life*. University of California Press, 2011.

[27] Dan Lockton, Devika Singh, Saloni Sabnis, Michelle Chou, Sarah Foley, and Alejandro Pantoja. New Metaphors: A Workshop Method for Generating Ideas and Reframing Problems in Design and Beyond. In *Proceedings of the 2019 on Creativity and Cognition*, pages 319–332, New York, NY, USA, 2019. ACM.

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

[29] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2), 2019.

[30] Dan P McAdams, Michelle Albaugh, Emily Farber, Jennifer Daniels, Regina L Logan, and Brad Olson. Family metaphors and moral intuitions: How conservatives and liberals narrate their lives. *Journal of personality and social psychology*, 95(4):978, 2008.

[31] Karin S Moser. Metaphors as symbolic environment of the self: How self-knowledge is expressed verbally. *Current Research in Social Psychology*, 12:151–178, 2007.

[32] Andreas Musolff. Metaphor and cultural cognition. In *Advances in Cultural Linguistics*, pages 325–344. Springer, 2017.

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

[34] Sarah T Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, CT, USA, 2019.

[35] Minna Ruckenstein and Linda Lisa Maria Turunen. Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, 22(6):1026–1042, 2020.

[36] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. It takes a village: Integrating an adaptive chatbot into an online gaming community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, 2020. ACM.

[37] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[38] 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.

[39] 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, New York, NY, USA, 1st edition, 1999.

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

[41] Donghee Yvette Wohn. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 160:1–160:13, New York, NY, USA, 2019. ACM.

[42] Bingjie Yu, Katta Spiel, Joseph Seering, and Leon Watts. "Taking Care of a Fruit Tree": Nurturing as a Layer of Concern in Online Community Moderation. In *CHI '20 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '20, pages 253:1–9, New York, NY, USA, 2020. ACM.

5

Linguistic Factors in Rebukes and Rules

5.1 Introduction[1]

> *"I'm white and I have zero complaints about how I'm treated by anyone because I'm not a confused little pussy like you."* – Reddit user

> *"I didn't think I would have to even put it in the side bar but I am not going to allow the sale of puppies and/or dogs on this subreddit. This is a discussion forum, not a place to hawk your wares."* – (/r/Rottweilier)

A core part of norm-setting in social interactions is the choices we make about what language to use. For example, we use *rebukes* to indicate when a person's behavior is unacceptable, but rebukes can take a wide variety of forms; they vary from detailed explanations of how a person has erred to profanity-laced diatribes. Similarly, *rules* are the codification of values and expectations, and can be written in many different ways. A rule could be clear and to the point, e.g., "From now on all post [sic] containing animated .gifs or high-speed videos will be removed." (/r/TimeLapsePorn), or full of value-laden language as in the quote above about selling puppies or dogs.

[1]This work was done with significant support from Diyi Yang and Tony Wang

Much prior social computing research has explored the impact of norms, whether by rule-setting, example-setting, or reward and punishment mechanisms. However, relatively little of this work has explicitly explored the linguistic properties of these behaviors, especially in the context of rebukes. These actions form the basis of the in-the-moment processes used by moderators [18]; though moderators may not think of rebukes as technically a moderation tool, extensive prior work has shown the importance of moderators' social behaviors [17, 6]. Where Chapter Two of this work examined content moderation from a macro lens, focusing on platform-sized ecosystems, and Chapter Three studied moderators' self-concepts and social roles over time from more of a meso lens, this chapter takes a micro approach to content moderation.

In this chapter, I present results from studies of both rebukes and rules. In each case, I performed a study on data from Reddit prior to running an experiment on a simulated discussion forum via Mechanical Turk, informed by the results of the first study. Though these studies are preliminary, they provide solid evidence and suggestions for promising directions in future research in these areas. The studies I present here are, in the order they appear in the text:

1. **Reddit Rebukes:** What impact does "rebuking" another user have on their subsequent behavior and the behaviors of others participating in the discussion?

2. **Discussion forum Rebukes:** How does using different types of personal and social identity language impact how effective a rebuke is?

3. **Reddit Rules:** What happens after a rule change in a subreddit is announced, and what factors affect how this plays out?

4. **Discussion forum Rules:** How does using different types of personal and social identity language impact how effective a rule is?

5.2 Related work

Though there are many frameworks that could be applied to the analysis of norm-enforcing language, I have chosen to focus here on properties of comments including analytical complexity and anger, as these are two axes upon which Reddit comments vary significantly. Related work in this domain has looked at the impact of a variety of different factors. Munger found that rebukes on Twitter could reduce subsequent racist behavior of the target if the rebuker was both high-status and a part of the ingroup (white-presenting) [11]. He used bots to respond to tweets with the slur, "n****r" with a gentle but clear rebuke –

> "@[subject] Hey man, just remember that there are real people who arehurt [sic] when you harass them with that kind of language"

Munger drew from both social referents theory, which outlines the influence of high-status individuals [12], and the Social Identity Model of Deindividuation Effects [15], finding that anonymous individuals were more susceptible to rebukes from high-status ingroup members.

Other work has looked at persuasiveness in text on sites like Reddit. Tan et al. [21] find linguistic cues in both an expressed opinion and the response that help predict whether the response will successfully persuade its target. Both Yang and Kraut [22] and Durmus and Cardie [4] find impact of cues related to identity. Yang and Kraut find that cues that allowed lenders to identify with people soliciting loans made the lenders more likely to lend. Durmus and Cardie found that identity-related language made persuaders more likely to continue persuading in a debate.

In the first study here we use methods generally similar to those in Tan et al., looking at a comment on Reddit and a rebuke reply in order to assess whether behavior change occurred in response to the rebuke. Though, unlike in the /r/ChangeMyView community analyzed by Tan et al., users in the general Reddit space do not typically explicitly acknowledge when their opinions or behaviors have changed, we use a variant of sentiment analysis to look

at these users' subsequent behaviors and also analyze number of subsequent comments and their length. The second study transitions this work into an experimental forum, testing characteristics of rebukes to determine their impact on subsequent commenters.

Rules have also been studied at some length, though recent work has focused primarily on the importance of their display [8, 10]. In my prior work [18], I found that rule changes were typically preceded by a change in the community; moderators observed a behavior that they had not expected, and had to figure out if and how to write a rule to respond to it. The third study here examines what happens when rules change in subreddits, and in the fourth study I performed an experiment to test how the language used in rules impacts users' subsequent comments.

5.3 Study 1: Impact of Rebukes on Reddit

In this work we have chosen to focus specifically on the act of rebuking rather than on a specific type of problematic behaviors. In this sense, as with Chandrasekharan et al.'s "Crossmod" [3, 1] I have not created a classifier that detects problematic behavior; instead, I have created a classifier that detects behavior that Reddit users think is problematic and, in the case of this study, worth rebuking. The detection of rebukes, though a complex linguistic task, offers a number of different challenges. While hate speech and harassment suffer from a glut of definitions across literature from different fields, the term "rebuke" has been used only very rarely. We base our analysis here on a definition from Radzik's "Moral Rebukes and Social Avoidance":

> *"A rebuke is a pointed expression of disapproval—usually but not necessarily moral disapproval—that is addressed to a perceived transgressor. A rebuke can take the form of a long diatribe, a single word spoken in private, a letter to the editor that is intended to be read by the public as well as the wrongdoer, or even an indignant glance across a room. It may use explicitly moral terms, foul lan-*

guage, or body language. 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. Criticisms can be gentle, loving and even wrapped in humor. Their overall tone can be one of sadness or disappointment. But moral rebukes are a form of criticism that expresses some degree of resentment or indignation. They convey anger even when other, competing emotions are also on display. Rebukes tell the wrongdoer, not just that she did something blameworthy, but that the speaker blames her. Rebukes do not just express the belief that she is responsible; they hold her responsible." [14, p. 644]

As Radzik's definition shows, rebukes come in a wide variety of forms, and can be couched in language that ranges from direct to vague and sarcastic. Here, as noted above, I focused specifically on the presence of analytical and anger-based language in rebukes. Note that, while the original plan for this project involved including whether the rebuking user was a moderator as a factor in the analyses, I found that only 32 out of nearly 100,000 rebukes in my final dataset were made by moderators. This shows that, while moderators may respond to posts to explain why behavior is inappropriate [6], they rarely do so in a way that could be classified as a rebuke.

In this study I focused primarily on groups of three sequential comments: An **initial comment** made by one user, a **reply** by a second user, which *may or may not be a rebuke*, and a **follow-up response** by the first user. I also analyzed comments later down the same comment trees when the tree continued past the first three comments.

This study asks the following questions:

1. How does a rebuke impact the subsequent behavior of the rebuked person?

2. How does the presence of analytical and anger-based language in the rebuke moderate this effect?

3. How does this rebuke impact subsequent participation in the conversation by other users?

5.3.1 Data collection and model building

In order to answer these questions, I first needed to build a classifier to accurately identify rebukes at scale. In order to do this, I scraped comments from the past month from twenty different subreddits. I selected ten subreddits that had been identified in previous research to be conflict-heavy [2, 9], and ten corresponding low-conflict subreddits of similar sizes in order to have a balanced dataset. I led a team to manually sort through these comments to identify rebukes of any sort, per the definition above. We were able to gather approximately 500 examples of rebukes, from which I constructed a basic classifier to support faster searching. With the help of this classifier, we were able to find 1055 total examples of rebukes, and we matched these rebukes with 1055 non-rebuke comments.

From these comments, I developed a more advanced classifier with the goal of differentiating comments at scale in order to be able to build statistical models. I used an SVC (Support Vector Machine for classification) model based on textual properties of comments classified as rebukes vs non-rebukes, using a term-frequency inverse document frequency (tf-idf) transformation and a grid search for hyperparameter tuning. On a balanced test set, this model achieved a 73.93% accuracy score with 0.479 Kappa. Subsequent analysis showed that adding more data to the training set only marginally improved performance.

While this performance was significantly better than random, I elected to set additional prediction threshholds. As the goal in this project was not to accurately classify every comment but rather to identify differences between the impact of rebukes and non-rebukes, I elected to select comments where the model had predicted one outcome with high confidence. After qualitative examination of the predictions, I set these threshholds at < 0.15 for non-rebukes and > 0.65 for rebukes. These threshholds led to low false positive and false negative rates.

In order to be able to match similar comments, I also created a "rebukable" classifier, which predicted whether a given comment would be rebuked. Like the previous classifier, this classifier was an SVC model based on textual properties of comments that were known to have been rebuked vs not rebuked, using a term-frequency inverse document frequency (tf-idf) transformation. Because of the similarity between this dataset and the previous dataset, and the lack of value previously gained from a grid search for hyperparameter tuning, a grid search was not performed to optimize this model. This model achieved a 60.07% accuracy score with 0.203 Kappa. These scores were lower than scores for the previous model, but this was expected; the previous model predicted only the intent behind a given piece of text (i.e., whether it was intended as a rebuke), while this model predicted how a different person would react to the meaning of a given piece of text (i.e., whether they would rebuke or not rebuke), a more socially complex task.

Figures 5.3.1 and 5.3.1 show characteristics of both of these models. Figure 5.3.1 shows the terms most strongly associated with whether a comment would be classified as a rebuke or not a rebuke. These terms were selected from among the top 100 most common words in the dataset after stopwords had been removed, and include the words that had the highest and lowest weights in the SVC model.

More likely to be a rebuke

	Stem	Comments	Coeff
1.	moron	56	0.96
2.	ignor	52	0.76
3.	a**hol	47	0.74
4.	idiot	74	0.68
5.	b**ch	47	0.65
6.	murder	64	0.63
7.	racist	69	0.62
8.	comment	130	0.61
9.	does	52	0.57
10.	stop	75	0.55
11.	talk	62	0.51
12.	f**k	331	0.5
13.	wrong	54	0.48
14.	men	70	0.47
15.	let	65	0.45

More likely NOT to be a rebuke

	Stem	Comments	Coeff
1.	come	54	-0.25
2.	sure	67	-0.26
3.	guy	83	-0.27
4.	want	119	-0.3
5.	day	50	-0.32
6.	isn	60	-0.32
7.	ll (I'll, we'll)	53	-0.39
8.	mean	77	-0.41
9.	love	49	-0.43
10.	feel	79	-0.43
11.	lot	48	-0.44
12.	didn	57	-0.44
13.	start	55	-0.51
14.	Use	92	-0.62
15.	yes	47	-0.81

Figure 5-1: Terms most strongly associated with whether a comment was predicted to be a rebuke or not a rebuke.

More likely to be rebuked

	Stem	Comments	Coeff
1.	trump	81	2.06
2.	women	68	1.69
3.	care	66	1.49
4.	hate	60	1.41
5.	sh*t	95	1.38
6.	girl	81	1.32
7.	guy	96	1.28
8.	look	105	1.25
9.	f**k	141	1.17
10.	understand	53	1.12
11.	turn	67	1.12
12.	away	71	1.06
13.	post	70	1.03
14.	don	266	1.00
15.	peopl	264	0.98

More likely NOT to be rebuked

	Stem	Comments	Coeff
1.	right	120	-0.93
2.	relationship	56	-0.93
3.	game	55	-0.95
4.	question	56	-0.96
5.	doesn	95	-1.00
6.	watch	63	-1.09
7.	good	135	-1.12
8.	go	109	-1.13
9.	end	68	-1.16
10.	place	59	-1.25
11.	want	200	-1.31
12.	day	126	-1.36
13.	need	137	-1.62
14.	year	142	-1.73
15.	love	84	-1.84

Figure 5-2: Terms most strongly associated with whether a comment was predicted to be rebuked or not rebuked.

Many words on these lists match expectations for characteristics of the associated comment types. For example, many of the terms most associated with a comment being classified as a rebuke were curse words (e.g., a**hole, f**k) or insults (moron, racist, idiot). Similarly, terms most associated with comments predicted to have been rebuked were associated with emotions (care, hate, understand), curses (sh*t, f**k), or contentious issues (trump, women, girl, guy).[2] Words associated with predicted non-rebukes and non-rebuked comments were fairly neutral or sometimes positive (love, good, right, yes).

Following completion of these classifiers, I used the Pushshift API to gather 160 million Reddit comments from late 2019. I selected a subset of these comments to again reflect a balance of conflict-heavy and non-conflict-heavy subreddits. I assembled these comments into trees in JSON format, and parsed each full branch of each tree to identify branches that began with the format described above – an initial comment, a reply from another user that might or might not have been a rebuke, and a follow-up response from the first user – and stored each of these branches separately for analysis. Finally, I removed all branches from this dataset where any of the three comments had been deleted. The result of this down-selection was a dataset of 94,442 sequences of comments matching this format.

[2]The dataset included subreddits that frequently discussed controversial gender issues.

I then used scripts to create a list of characteristics of each branch. These included the following:

- Analytic score, Anger score,[3] and length of the first, second, and third comments.
- The number of additional comments in the branch.
- How strongly the second classifier predicted that the first comment would or would not be rebuked.
- Whether the first classifier had identified the second comment (the reply) as a rebuke.

From these characteristics I also created three additional characteristics: the change in Analytic and Anger scores and the change in length between the initial comment and the follow-up reply after the rebuke or non-rebuke. I normalized all of these characteristics around a mean of 0 and standard deviation of 1. In total, 44,367 of the reply comments were classified as non-rebukes and 50,075 were classified as rebukes.

5.3.2 Results

In order to answer the research questions in this study, I performed two types of analyses. First, I compared branches with rebuke replies to branches with non-rebuke replies in order to isolate the impact of the rebukes. Next, I focused only on branches with rebuke replies to identify what factors moderate the impact of these rebukes.

Analysis One

In the first analysis I compare dependent variables based on whether the first reply was classified as a rebuke. I created pairs of comment trees where one tree included a rebuke and the other did not, but where the second classifier (whether the original comment would receive a rebuke) had given a near identical score for both comments. This allowed me to

[3] Analytic and Anger scores were calculated using dictionaries from the LIWC 2015 toolkit [13]

separate the effect of the rebuke from the effects of the relevant characteristics of the initial comment. I used a paired-samples t-test with dependent variables including normalized change in *Analytic score* and *Anger score* between the original post and the same user's follow-up reply after the rebuke (Analytic or Anger of the third minus Analytic or Anger of the first comment in the thread); normalized change in *length* in total characters between the original post and the same user's follow-up reply after the rebuke; and the normalized number of additional comments in the branch after the follow-up reply.

	Mean additional increase after rebuke	t	p	Cohen's d
Normalized length	0.18 (1.3)	20.40	<0.001***	1.32
Normalized Analytic score	0.18 (1.4)	19.45	<0.001***	1.40
Normalized Anger score	0.01 (1.4)	1.59	0.110	1.38
Normalized later comments	0.78 (1.4)	83.39	<0.001***	1.41

Table 5.1: Paired t-test comparisons when the first reply was classified as a rebuke vs not classified as a rebuke. Pairs had matched scores from likeliness to have been rebuked classifier. Total pairs = 22,572

Table 5.1 shows that the presence of a rebuke significantly impacted several of these dependent variables. Follow-up replies from the original poster were shorter than the original comment in both cases, but after a rebuke the follow-up reply was longer than after a non-rebuke reply. Follow-up replies were also more analytical after a rebuke than after a non-rebuke. Surprisingly, the presence of a rebuke did not significantly impact the change in anger score between the original comment and the follow-up response to the rebuke. After a rebuke, there were significantly more comments later in the branch.

These results suggest that rebukes sparked additional conversation; the original commenter wrote more and in more depth, perhaps to explain or justify their position, and more users chimed in to add their own opinions.

Analysis Two

The second analysis explored whether the textual properties of rebukes identified earlier, Analytic score and Anger score, impacted subsequent outcomes. For these analyses, the

branches that did not contain rebukes were removed in order to compare only between different types of rebukes. I examined the same four dependent variables: change in length of response to the rebuke, change in Analytic score, change in Anger score, and change in number of later comments (all normalized). I ran four linear regressions, each with one of these as the dependent variable, and the independent variables being properties of the rebuke including normalized Analytic score and normalized Anger score.

Tables 5.2 and 5.3 show the linear regression models for change in length and anger of response to the rebuke. The other two models, which showed impact on analytic score and number of later comments, did not show significant impact of either independent variable so they are excluded here. Though coefficients for all significant effects were small, these tables show that increased anger in the rebuke unsurprisingly leads to increased anger in the subsequent response, but that increased analytic score of the rebuke does not lead to changed analytic score of the subsequent response; instead, it leads to decreased length of the response. One plausible explanation for this is that a deeply analytical rebuke could discourage the initial commenter from responding in depth because they might have to rebut a reasoned argument or because they realize that their original argument lacked substance relative to the rebuke. Further exploration is needed to gain a more nuanced understanding of these and related effects.

It is also plausible that these effects appear in other places. While this study focused on trios of comments where the author of the first comment was also the author of the third, it is possible that other users replied directly to the rebuke to voice their agreement or opposition. However, this analysis would fall under a more general analysis of conflict in online discussions rather than an analysis of rebukes, and as such it is left for future exploration.

Table 5.2: Linear regression for normalized Δ anger of response

Factor	Coefficient	StDev	t	p
(Intercept)	0.000	0.007	0.000	1.000
Normalized Analytic$_{rebuke}$	0.010	0.007	1.555	0.120
Normalized Anger$_{rebuke}$	0.013	0.007	1.961	0.049*

Table 5.3: Linear regression for normalized Δ length of response

Factor	Coefficient	StDev	t	p
(Intercept)	0.000	0.007	0.000	1.000
Normalized Analytic$_{rebuke}$	-0.015	0.007	-2.202	0.028*
Normalized Anger$_{rebuke}$	-0.001	0.007	-0.785	0.432

5.4 Study 2: Rebukes in a controlled comment thread

In the next study, I chose to explore whether the use of identity-based language in a rebuke could impact subsequent comments. I used the fake political blog post format that I have used in previous studies, e.g., [16]. In this structure, MTurk workers are exposed to a fake political blog post with fake comments listed below, and are asked to contribute their own comment.[4] After submitting their comment, MTurk workers filled out a brief survey to note their age, gender, and political leaning on a seven point scale.

While in previous studies where I used this format there were multiple comments shown beneath the blog post, my goal with this study was to isolate the impact of a single comment, as in the previous analysis of rebukes on Reddit. Figure 5.4 shows the setup of this experiment. Note that, unlike the previous Reddit study, I could not examine the impact of a rebuke on an individual's follow-up comment because of the format of the experiment. This experiment instead looks at the impact of observing a rebuke on a different person's subsequent behavior.

This experiment contained two variables and seven conditions. Each variable was a manifestation of two different facets of identity. The first varible was the use of pronouns in

[4]Workers were paid $1 for a 5 minute study, averaging to a rate of $12 per hour.

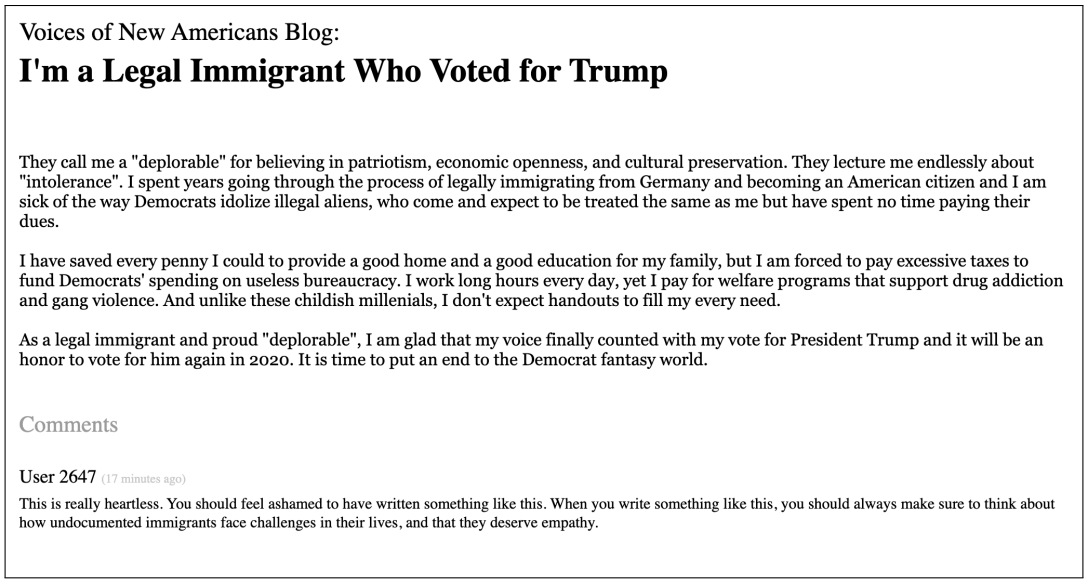


Figure 5-3: Final version of blog post with rebuke

the rebuke to refer to the author of the blog post, the author of the rebuke, or a general 'we', e.g., "[I would]/[You should]/[We should] feel ashamed". The second identity facet referred to the social group being discussed in the blog post. Either the social group was referred to in a general sense or in a specific sense – "you should always make sure to think about how **all different types of people** face challenges in their lives, and that they deserve empathy" or "you should always make sure to think about how **undocumented immigrants** face challenges in their lives, and that they deserve empathy". This variable is referred to as "Diversity" in the models below, and is intended in some part to mirror the current public conversation about "All lives matter" and "Black lives matter". The experiment also contained a control condition where users were not shown a rebuke.

As in the previous version of this study, the dependent variables were the Analytic, Tone, and Social scores of the comments that MTurk workers posted after having read the post and the rebuke, as assigned by the LIWC statistical package [13]. The Analytic score rates the sophistication of the wording used in the comment, focusing on logical words; the Tone score rates positive vs negative tone; and the Social score quantifies use of words describing social interconnectedness. Building from the previous study, I also included word count as a

	Analytic	Tone	Social	WC
No rebuke comment	50.3 (32.8)	59.5 (38.9)	13.2 (7.3)	44.7 (39.7)
1^{st} p sing * All Lives	34.9 (29.3)	56.6 (35.7)	13.3 (7.0)	53.0 (48.1)
1^{st} p sing * Undoc. Imm. Lives	42.5 (33.6)	55.5 (38.8)	14.0 (8.5)	36.3 (29.2)
2^{nd} p sing * All Lives	43.3 (30.0)	48.5 (40.1)	13.8 (6.5)	45.8 (27.8)
2^{nd} p sing * Undoc. Imm. Lives	52.5 (29.6)	65.6 (39.0)	13.5 (7.3)	52.5 (29.6)
1^{st} p plural * All Lives	52.8 (28.1)	49.0 (42.0)	12.0 (5.6)	41.7 (22.7)
1^{st} p plural * Undoc. Imm. Lives	46.1 (31.3)	59.5 (39.0)	11.8 (5.2)	50.3 (27.4)

Table 5.4: Means and standard deviations for Analytic, Tone, and Social scores of comments and their Word Counts (WC) as assigned by LIWC

Dependent Variable: Analytic			
	Mean Square	F	p
Corrected Model	1710.26	1.83	0.06
Intercept	12586.09	13.47	0.00
Politics	72.62	0.078	0.78
Age	4252.55	4.551	0.03*
Gender_Int	2977.03	3.186	0.08'
Diversity	810.44	0.867	0.35
Pronoun	2500.81	2.676	0.07'
Diversity * Pronoun	719.57	0.77	0.46
Error	934.38		
R Squared = .065 (Adjusted R Squared = .030)			

dependent variable. Table 5.4 shows the means and standard deviations for each condition in each of these dependent variables.

In order to determine whether these two factors impacted the comments subsequently posted by workers, I performed univariate ANCOVA analyses with diversity identity language (all lives vs undocumented immigrant lives) and pronouns as fixed effects and politics, age, and gender as covariates.[5] I have elected not to include all models here; most terms did not have significant impact on the outcomes as measured. I include two models where marginally significant effects were found for the interventions.

These effects are not statistically significant at $p < 0.05$, but they do provide some evidence that variants on these interventions may have potential. In the analysis for the Analytic dependent variable, the first person singular pronoun led to the lowest Analytic

[5]Note that Gender_Int is a variable where gender was converted to a binary variable for the purpose of making it a covariate.

Dependent Variable: Word Count			
	Mean Square	F	p
Corrected Model	2516.42	2.22	0.02
Intercept	16721.87	14.75	0.00
Politics	7933.60	7.00	0.01**
Age	4796.21	4.231	0.04*
Gender_Int	2877.82	2.539	0.11
Diversity	181.02	0.16	0.69
Pronoun	77.63	0.068	0.93
Diversity * Pronoun	2852.87	2.52	0.08'
Error	1133.59		
R Squared = .078 (Adjusted R Squared = .043)			

scores for responses, and in the Word Count analysis, the first person singular * undocumented immigrant lives condition had a lower word count than the other conditions. This might be a result of the distancing that could occur in these conditions – both the second person and first person plural pronouns directly engage in an interpersonal way, while the first person singular pronoun does not explicitly extend the social range of the rebuke beyond its author. Similarly, the Undocumented Immigrant Lives condition, in combination with this first person singular pronoun, may have just been too distant from most participants' experiences to evoke a substantive response. Further work is necessary to provide concrete explanations for the trends seen here and to evoke more concrete results.

5.5 Study 3: Rules on Reddit

The third analysis in this set looked at what happens when rule changes are made on Reddit. While prior work has examined and discussed the impact of displaying rules [8, 10], no work has explicitly attempted to quantify the immediate impact of changing rules across different online communities.

5.5.1 Data collection and analysis

In order to identify rules changes, I manually sorted through a subset of 2000 posts from a corpus of 7746 posts gathered from Reddit via the Pushshift API that mentioned rules and/or rule changes. While I had originally intended to focus on rules that governed behavior in comments, I quickly realized that rules governing the content of posts were far more common in the dataset.

I used a number of criteria to identify candidate rule posts. The following posts were excluded from my dataset:

1. Rule changes from subreddits that have been banned
2. Rule changes from subreddits that have been closed by their moderators
3. Rule changes from subreddits that are invite only
4. Rule changes where the post had been deleted or removed
5. Rule changes where the post was just a pointer to the sidebar
6. Rule changes that were listed in external documents
7. Rule changes from subreddits with less than 1k subscribers, as the volume of behavior would likely be too small to analyze
8. Posts that were reminders to follow the rules
9. Posts that were introductions to the rules for the first time
10. Posts asking for suggestions for rules
11. Posts asking for feedback about rule changes
12. Posts where rules were removed rather than added

13. Rule changes in subreddits that are designed to host on-Reddit role-playing games, when the rule changes were about in-game rules

This left a relatively small subset of posts from within the 2000 posts I reviewed. I chose 30 posts, 15 of which described "functional" rule changes and 15 of which described "value-based" rule changes. Functional rule changes included general requirements for post format or frequency, e.g., "Only one photograph is allowed per submission" (/r/photocritique) or "Once you've made a post, please wait a minimum of 7 days before posting again" (/r/makeupexchange). Value-based rule changes describe a type of content or behavior that the moderators have elected to exclude for generally moral reasons, e.g., "Please refrain from adding opinion to the reddit subjects, when linking to a news item... Please also try and avoid rumour. Instead try and track down sources for the information you want to present" (/r/StLouis) or "Posts asking for or providing pirated materials will be deleted." (/r/medicalschool). This division, between logistical and values-based rules, emerged as qualitative analysis of post text proceeded. These two distinct categories emerged consistently and reliably as different types of rules, so it made sense to differentiate them in the analysis. The full list of rule changes in the dataset is as follows:

The primary dependent variables in this study were how many posts were made immediately before the rule changed as compared with immediately after, and correspondingly how many were deleted in each time range. I elected to use a time range of one week before and one week after the date of the rule change post. I collected all posts from each subreddit in the above lists within this range, calculated as post time in UTC minus 604800 to post time plus 604800, where 604800 equals one week in UTC. These posts and deletions were separated into bins by day, periods of 86400 (seconds) in UTC. For each rule change, the resulting dataset contained the full text of the rule post, the name of the subreddit, whether the rule change was logistical or values-based, how many posts were made and how many were deleted in each day in the 14 day range, and calculated variables showing the difference between each of the total and deletions variables from before to after the rule change

Table 5.5: **Functional** rule changes

Subreddit	Rule summary
Diablo	Excessive posting of the same information in a short period of time will result in all posts removed except for the most active post
AsianNSFW	Imgur is now a requirement for all pictures
HeroesandGenerals	No self-promotion by linking products/social media pages/ streaming services.
photocritique	Only one photograph is allowed per submission
nononomaybe	Posts must be in 50/50 format or it must have two possible outcomes in the title. The two outcomes in the title must be related to one other.
ShitCosmoSays	Please add a NSFW [tag] if there's a sex position [diagram]
PrettyLittleLiars	No more blog links.
TimeLapsePorn	From now on all post containing animated .gifs or high-speed videos will be removed.
SchoolIdolFestival	Do not post spoilers from the anime without a spoiler tag, unless in a discussion thread already tagged with [SPOILERS]
AsianP****	Any Gravure.com content is no longer allowed in submissions.
reactiongifs	If a submission is a repost of a submission in the top 100 of all time it will be removed.
makeupexchange	Once you've made a post, please wait a minimum of 7 days before posting again
smashbros34	There's been plenty of posts without proper titles [...] Just type the character's name so it will be much easier to search up for others.
kinky	All images must be direct links to a reputable image host.
Minecraft	Posts must contain minecraft- related content in the link/post body, not just title.

Table 5.6: **Values-based** rule changes

Subreddit	Rule summary
HongKong	No Editorialized Titles
ChristianGirls	Thou shalt not post any suggestive or sexual content featuring minors. (Seriously, Hell will be the least of your worries.)
GoneWildPlus	This is a haven for appreciation. If you're not into a given individual, you shouldn't feel the need to tell them.
Tgirls	No more escort posts.
Rottweiler	I am not going to allow the sale of puppies and/or dogs on this subreddit. This is a discussion forum, not a place to hawk your wares.
evangelion	All illegal seeking or supplying of copyrighted material is banned from the subreddit.
medicalschool	Posts asking for or providing pirated materials will be deleted.
conspiratard	No submissions of personal spats, please
wow	For every one link to content you created, be it artwork, an external guide, etc, you should submit ten other links or comments to r/wow.
fireemblem	no shit posting
wtfamazon	items must be "weird" of some sort
WeddingPhotography	No self posting of website/blog article links
RWBY	Submission of bootlegged, filmed, stolen or any unofficial RWBY episodes to this subreddit will be removed immediately
StLouis	Please refrain from adding opinion to the reddit subjects, when linking to a news item. [...] Please also try and avoid rumour.
ABraThatFits	I am going to implement a "No cussing during AMAs" rule

(pre-deletions minus post-deletions; pre-posts minus post-posts). "Clout" and "Social" scores from LIWC were added for each rule change based on the text of the post. Each of these values was then normalized by subtracting the mean and dividing by the standard deviation.

"Clout" scores represent a measure of the authoritativeness of the post; authority is a core part of the social identity theory of leadership [5], and this theory predicts that expressions of authority by an established leader will have greater impact than those from a non-leader. As all of these posts were made by moderators, all were leaders at least as much as platform features allow. Thus, instead of measuring legitimacy of leaders, we instead measured the magnitude of the expression of authority with the hypothesis that it would have a similar effect. We also measured "Social" language via LIWC to capture the direct intragroup language with the hypothesis that increased use of intragroup language by a leader will increase compliance to the stated new rule.

Initial exploration of the data did not prove promising for finding explanatory relationships; a paired samples t-test found insufficient evidence to reject the null hypothesis that the pre-deletions and post-deletions distributions were the same ($p = 0.43$) and the results were similar for the pre-posts and post-posts distributions ($p = 0.69$).

However, in order to be thorough, I performed additional linear regressions modeling the differences bewteen normalized pre-deletions and normalized post-deletions and, correspondingly, the differences between normalized pre-posts and normalized post-posts. The dependent variables in these regressions were Clout score, Social score, and rule change type (0 = logistical, 1 = values-based). In the regression modeling the difference in normalized total posts before and after the rule change, none of these factors were found to be significant. In the regression modeling the difference in normalized total deletions before and after the rule change, rule change type was found to be marginally significant. As shown in Table 5.5.1, values-based rule changes led to slightly fewer subsequent deletions than logistical rule changes. This is a minor difference, totaling one third of a standard deviation. This provides some justification for making a distinction between these two types of rules in future work.

Linear regression for normalized difference in deletions			
	Standardized B	t	p
(Constant)		-0.97	0.34
Type_Numeric	-0.33	-1.84	0.08’
Clout score	0.16	0.76	0.46
Social score	0.14	0.65	0.52

Neither the Clout score nor the Social score, as calculated by LIWC 2015 [13] had any significant impact on either dependent variable. This suggests at minimum three possible future hypotheses – first, that language of rule posts does not significantly impact subsequent posting behaviors; second, that alternative linguistic features have more impact than those tested here; and third, that the features tested here do have a significant impact, but a larger sample size is required to detect it.

Though the quantitative results from this study do not provide solid evidence to support the hypotheses that authority (Clout) or social language (Social) impacts subsequent behaviors, a brief qualitative analysis of the collected rules does provide some interesting points that both confirm prior work and suggest directions for follow-up work.

The emergence of new rules

Prior work [18, 20] found that rules are frequently created in response to unexpected changes in the community; moderators observed a new behavior that they had not anticipated, or saw an increase of an existing behavior to the point where it became disruptive. This was borne out in a large fraction of the collected rules:

> *“I have noticed an increase of people linking to their own blogs with their own theories of things relating to the show. While it’s great that you took the time to write it out and want people to read it, here is not the place.*
>
> [...]
>
> *There have been quite a few repeat offenders, but since there was no rule against it I haven’t done anything.”* – /r/PrettyLittleLiars moderator

One rule change in the dataset reflected a very unusual historic situation: the fatal shooting of Michael Brown in Ferguson, Missouri in August 2014.

> *It's now been a week since the shooting in Ferguson. We've had a lot of threads here as various bits of news have come out. We also have a great continuing live feed of the news, scanner transcriptions, tweets, and other sources about this situation. The guys contributing to that have done a great job in keeping many people informed.*
>
> *As you can see from our traffic stats, we've had roughly 5 times the number of unique visitors per day than normal, and we've had roughly 10 times the number of page views per day. We've also dramatically increased our subscribership.*
>
> [...]
>
> *Please refrain from adding opinion to the reddit subjects, when linking to a news item. Instead give a subject that reflects the title of the piece (if in doubt, copying the title directly is usually a good plan). Feel free to express your opinions in the comments though.*
>
> *Please also try and avoid rumour. Instead try and track down sources for the information you want to present. reddiquette says "Please do Look for the original source of content, and submit that." If a tweet references a news article, link to that news article rather than the tweet.* – /r/StLouis moderator

In the former quote, an increase in a particular type of unwanted behavior caused a moderator to make it clear that they would begin taking action against that type of behavior. In the latter quote, an external social situation had a huge impact on the subreddit community, leading the moderation team to clarify norms and establish additional standards. Both of these scenarios are in line with prior work, e.g., [7, 18, 19] and also the examples given in Chapter three of this work (e.g., the Colin Kaepernick example).

A very basic but likely useful research question to explore building from this analysis would be to ask, "what are the different types of situations that lead to different types of rules changes?"

Rules and NSFW spaces

A significant fraction of the new rule posts, rule change posts, and rule reminder posts came from NSFW (not safe for work) subreddits, which typically contain nudity or other adult content. One might think that these subreddits are especially important to moderate carefully because of the somewhat-taboo nature of their content, and in some cases this was true, as in the first rule in the rule change post from /r/ChristianGirls:

> *"Thou shalt not post any suggestive or sexual content featuring minors. (Seriously, Hell will be the least of your worries.)"* – /r/ChristianGirls moderator

This rule is very clearly important both for moral and legal reasons. However, the remaining rule changes from this post are significantly less serious in their topics:

> *"Thou shalt only post images that are NSFW, so sayeth the Lord. This is not the place to post church pictures, it's the place to post pictures that would get you banned from the Church.*
>
> *If thou wishes to share a video, it must be done in the comments of the post, not as a post itself. Posts that are not RES-previewable will be removed henceforth.*
>
> *Thou shalt not be a dick, especially in regards to self-posters. We are fortunate to have ladies that post pictures of themselves here and we want to create a friendly and welcoming atmosphere to encourage repeat visits."* – /r/ChristianGirls moderator

Beyond their comical and thematically appropriate wording, these rules are more about creating the desired type of community both through values and logistics.

Other subreddits had even more mundane rule changes. From /r/AsianP****:

> *"It has now become clear that the Gravure.com watchdogs are taking down any and all of their content submitted to this subreddit, whether it be albums or individual pictures. Therefore I have no choice but to make an amendment to the recently added rule banning all Gravure.com content in this subreddit so as to avoid the subreddit getting littered with dead links."* – /r/AsianP**** moderator

and from /r/kinky: *"All images must be direct links to a reputable image host. Hopefully this will help cut back on spam"* and a similar rule from AsianNSFW: *"Imgur is now a requirement for all pictures. This will help cut down on spam, and 3rd party traffic."*

It would be valuable to further explore the development of rules in NSFW subreddits, which are historically under-researched spaces, in order to determine whether they face a different set of challenges or if they undergo the same general processes of governance identified in recent work [18].

5.6 Study 4: Rules in a controlled comment thread

Much as in Study 2, this study was a controlled experiment using the fake political blog post format [16]. In this version, however, comments were left out entirely. The only stimuli were the fake blog post and, when not in the control condition, a popup displaying a rule when the user clicked on the comment box to write their comment. Figure 5.6 shows this popup overlaid over the blog post and comment box, which are grayed out in the background until the user clicks "I AGREE".

While the previous study identified several threads for future exploration, the hypothesis that authoritative language would have a significant impact on subsequent behavior [5] was not supported by the analysis of the collected data. Therefore, I elected to run this experiment based on findings from the previous rebuke experiment, which found some evidence that identity-based language could infuence subsequent discussion.

In this study MTurk workers were again exposed to the fake political blog post, though

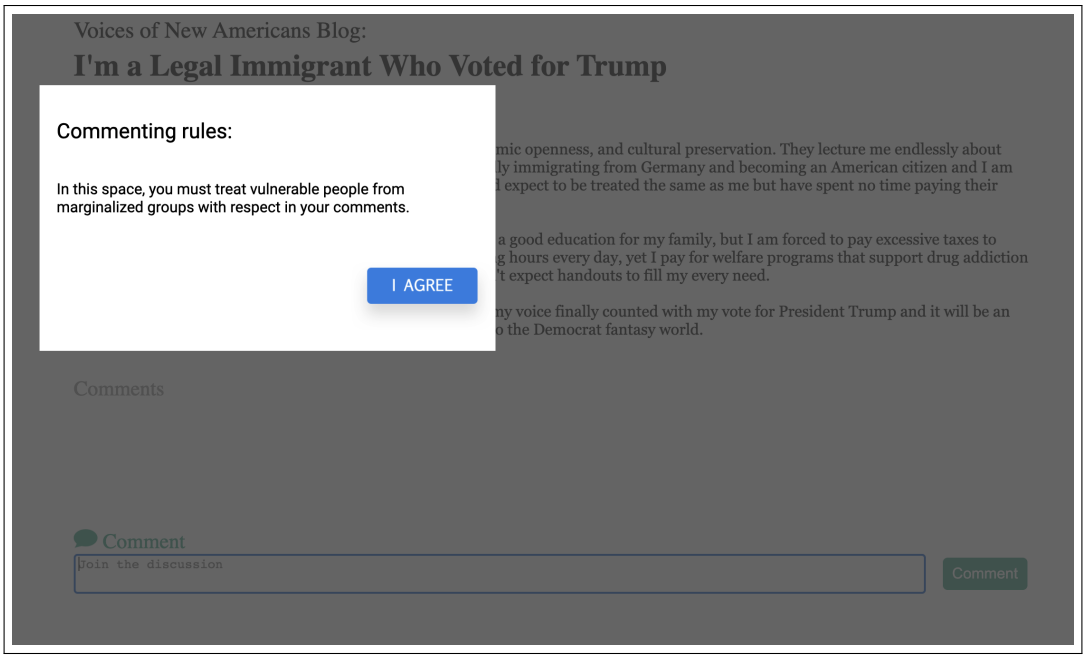


Figure 5-4: Final version of blog post with rebuke

without fake comments below, and were asked to contribute their own comment.[6] After submitting their comment, MTurk workers filled out a brief survey about their age, gender, and political leaning on a seven point scale.

This experiment contained two variables and five conditions. As in Study 2, each variable was a manifestation of two different facets of identity. The first varible was the use of pronouns in the rule shown on the popup, either "we" or "you" to refer to the participant. The second identity facet again referred to the social group being discussed in the blog post. Thus, the posible rule combinations were: In this space, [**we**/**you**] must treat [**all people**/**vulnerable people from marginalized groups**] with respect in [**our**/**your**] comments. The experiment also contained a control condition with no rules popup.

As in Study 2 and the previous version of this study, the dependent variables were the Analytic, Tone, and Social scores of the comments that MTurk workers posted, as well as their word count, as assigned by the LIWC statistical package [13]. Table 5.7 shows the means and standard deviations for each condition in each of these dependent variables.

[6]Workers were paid $1 for a 5 minute study, averaging to a rate of $12 per hour.

Condition	Analytic	Tone	Social	WC
No rules popup	53.6 (31.6)	58.3 (37.7)	11.7 (7.6)	43.3 (43.3)
1^{st} p plural * All people	50.5 (35.9)	59.1 (39.5)	12.0 (6.6)	39.0 (30.5)
1^{st} p plural * Vuln. marg. people	48.9 (28.8)	63.0 (37.4)	12.4 (6.3)	41.7 (27.9)
2^{nd} p sing * All people	50.0 (29.5)	63.1 (36.2)	11.3 (5.8)	59.7 (61.6)
2^{nd} p sing * Vuln. marg. people	55.4 (27.7)	57.1 (39.3)	12.1 (5.6)	49.5 (32.1)

Table 5.7: Means and standard deviations for Analytic, Tone, and Social scores of comments and their Word Counts (WC) as assigned by LIWC

Dependent Variable: Word Count			
	Mean Square	F	p
Corrected Model	8610.53	5.67	<0.001***
Intercept	13609.24	8.96	0.003**
Politics	11825.71	7.79	0.006**
Age	14990.78	9.87	0.002**
Gender_Int	24286.17	16.00	<0.001***
Diversity	825.80	0.54	0.462
Pronoun	9314.35	6.14	0.014*
Diversity * Pronoun	1503.56	0.99	0.321
Error	1518.35		
R Squared = .182 (Adjusted R Squared = .150)			

As in Study 2, I performed univariate ANCOVA analyses with diversity identity language (all people vs vulnerable marginalized people) and pronouns as fixed effects and politics, age, and gender as covariates, in order to determine what effect these factors had on the outcome measures.[7] As before, I have elected not to include all models here. I include one model, where the dependent variable was word count, where there were several significant effects. Note that politics was a significant factor in the "Tone" ANCOVA as well, but I do not include that model here because politics is a covariate and not significant for our theoretical questions.

As Table 5.6 shows, Pronoun significantly impacted the word count of participants' comments. The control group and first person plural groups had approximately the same word counts on average, but the second person singular groups had significantly higher word counts by approximately 35% on average. As with Study 2, it is plausible that this may have been

[7]Note that Gender_Int is a variable where gender was converted to a binary variable for the purpose of making it a covariate.

an effect of the social distance the participants felt between themselves and the rules. In the control condition, rules were not present, and in the first person plural condition rules were directed toward an unspecified "we" with whom the participant may not have identified. The second person singular condition directly targeted the participant. This possibility, however, does not directly answer the question of why this effect would have occurred only with word count. It is plausible that, when targeted by the direct, second person version of the rules, participants felt pressured to meet a certain standard, which led to an increase in volume of output but no change in the specific content. Future work could help to tease out these theoretical questions.

While the identity diversity condition (All people vs vulnerable, marginalized people) did not show a notable impact across the two comment thread experimental studies described in this chapter, there is evidence that they did have at least minor impact on some participants. For example, one participant (66 y.o, male, very conservative) wrote: "I can't really comment because I just learned that I cannot be disrespectful to the marginalized vulnerable, whoever they area, and I'm not sure who gets to decide what is respectful and what isn't." This effect clearly did not scale, but it is plausible that, given the right conditions, we might observe a reactance effect toward the use of terms like "marginalized and vulnerable people" among strong conservatives.

5.7 Conclusions

The above work provides evidence for potential future directions in exploring the textual nature of (non-moderator) interpersonal rebukes and the development of more effective rules. Though the findings do not directly provide explicit lessons for how to do either of these things more effectively, they do highlight the importance of considering the use of interpersonal, targeted identity-based language. This follows closely with predictions made by Social Identity Theory, which had not previously been tested in a linguistic analysis on the scale

or structure of the above studies.

More broadly, in the context of this dissertation, this work highlights that moderation happens on all different levels, from the organizational level discussed in Chapter 2, to the volunteer moderator level discussed in Chapter 3, down to the interactions between users discussed in this Chapter. In order to more fully understand the full ecosystem of moderation, all of these levels and their connections must be studied in more depth.

Bibliography

[1] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[2] 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. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):32:1–32:25, November 2018.

[3] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3175–3187, New York, NY, USA, 2017. ACM.

[4] Esin Durmus and Claire Cardie. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy, July 2019. Association for Computational Linguistics.

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

[6] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[7] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. Surviving an "eternal september": How an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1152–1156, New York, NY, USA, 2016. ACM.

[8] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. Regulating behavior in online communities. In Robert Kraut and Paul Resnick, editors, *Building Successful Online Communities: Evidence-Based Social Design*, chapter 4, pages 125–177. MIT Press, Cambridge, MA, USA, 2012.

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

[10] J Nathan Matias. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20):9785–9789, 2019.

[11] Kevin Munger. Tweetment Effects on the Tweeted : Experimentally Reducing Racist Harassment. *Political Behavior*, 39(3):629–649, 2016.

[12] Elizabeth Levy Paluck, Hana Shepherd, and Peter M Aronow. Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, 113(3):566–571, 2016.

[13] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[14] Linda Radzik. Moral Rebukes and Social Avoidance. *The Journal of Value Inquiry*, 48(4):643–661, dec 2014.

[15] 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.

[16] 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 606:1–606:14, New York, NY, USA, 2019. ACM.

[17] Joseph Seering, Robert Kraut, and Laura Dabbish. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 111–125, New York, NY, USA, 2017. ACM.

[18] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[19] 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, New York, NY, USA, 1st edition, 1999.

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

[21] 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*,

WWW '16, page 613–624, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[22] Diyi Yang and Robert E. Kraut. Persuading teammates to give: Systematic versus heuristic cues for soliciting loans. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December 2017.

6

Discussion and Directions for Future Work

In an article published in January of 1998, Johnson and Post argued that the future of internet governance was best served by a system where the full social web was divided into many groups, rather than governed centrally by national governments, with each group deciding on rules that mostly only impact the people within the groups.[1]

> *I think the answer to this important question is that a diverse set of rule spaces, coupled with real freedom of movement, structurally respects individual liberty (and minority opinions about values) to the greatest extent possible, even as compared with democratic top-down rule.* [20]

In essence, Johnson and Post proposed a system that looks something like a more purely democratic version of Facebook Groups, Reddit, or Discord, where a large portion of the users in any given community can participate in decision-making processess. If a community collectively acts seriously against the interest of a minority, these minority members can leave and form their own spaces. In extreme cases, the decision to expel a group (e.g., to ban a subreddit) can be made either by some large-scale democratic action or by a group of officials speaking for some sort of "sovereigns" (perhaps platform administrators or legislators)

[1] A modified version of this chapter combined with parts of the first and second chapters has been accepted to CSCW 2020 with Joseph Seering as solo author.

or a combination of both. This form of moderation is highly context-aware, as members of each group are much better situated to make decisions about their local communities than policymakers trying to make decisions for the whole system or even for each individual space.

Johnson and Post originally wrote this piece to argue for minimizing the extent to which existing territorial sovereigns (e.g., the German or American government) can regulate internet conduct, because they argued that these forms of governance would be less effective in meeting the twin goals of efficiency and legitimacy in governance. These geography-based arguments apply well in critique of Kaye's [21] proposed system of global social media governance. However, I suggest that Johnson and Post's proposal still fits well when "social media platforms" are substituted for "existing territorial sovereigns". Platform administrators, though not bound in the same way by geographic constraints, are frequently just as distant from the contexts of individual communities. If Johnson and Post's core argument is extended to say that centralized governance of any sort on the internet tends to be less efficient and less legitimate (per their definition of legitimacy), then it is reasonable to question the legitimacy and efficiency of the types of platform-driven online governance that have become so central to the modern web.

As discussed extensively above, a wide variety of popular, productive, and meaningful online communities have been run primarily through a system of self-moderation. These communities have emerged not only as stand-alone groups or on platforms that allow for self-moderation; some were formed in reaction to a for-profit platform's inability to moderate in a way that matched the community's norms. These communities show promise for the efficacy of community self-moderation as a model for content moderation, and more broadly for the possibility of a more user self-governed internet in the future. In this concluding chapter I discuss what this future might look like and how we can get there. I first provide a vision of how content moderation might work in the future, and I comment on some important pitfalls to avoid in working toward this vision. I next discuss what the path forward looks like for each level of content moderation – what needs to be done, what needs to be improved, and

what types of systems need to be built and tested. Finally, I present six guiding research questions that are intended to frame how academic researchers can help in moving toward a better, more user self-governed future for online content moderation.

6.1 A (Mildly) Radical Vision and Some Pitfalls

At the core of this dissertation has been the idea that user self-governance is a desirable thing, at least in moderation. This is an idea that is drawn directly from what I see as the core ethos of the field of Human-Computer Interaction – putting the human, the user, first. My perspective is fundamentally at odds with the Platforms and Policies perspective I described in Chapter 2, both because that perspective assumes that a top-down corporate and political model will be the de-facto future model but also because it tends to focus much more on the internal machinations of companies than on the experiences of users.

6.1.1 Radical Visions

In this section I articulate a vision for the future of content moderation that I believe is both achievable and user-centered. I focus here on structural elements of a future state. The following are several of these core elements, along with the social processes that occur within them:

Groups are at the heart of socialization

The default form for an online social space should be a *group*, rather than a *network*. As the previous chapters have shown both explicitly and implicitly, there are an extremely wide variety of options for governing groups, most notably including some forms of user self-governance. Social networks, on the other hand, are social structures that, by their very nature, are very difficult to self-govern because they lack the same bounded coherence. Thus, while groups are more amenable to multi-level governance (i.e., non-moderator users,

community moderators, platforms), networks are typically only amenable to governance by the first and third of these.

It is likely, though to my knowledge unproven, that, per social identity theory, interpersonal moderation strategies are less widely effective in networks because of the lack of structure brought by clear social identities; as many authors have discussed, social networks are prone to continuous context collapse, making it difficult to place oneself within a single identity and thus a single set of norms. For the above reasons, in this vision for the future the default (though not necessarily only) form for an online social space is a group.

Incentive structures push distributed rather than concentrated growth

As I note later in this chapter, the vast majority of online groups are very small to a point where they can reasonably be moderated by a small number of volunteers. In a future where groups are the default form for online social spaces, this means that the groups that will be most difficult to moderate will be the small number of extremely large groups, an assertion supported by work on current social spaces [19, 33, 34]. Thus, community self-governance will be more effective if incentive structures encourage the creation and maintenance of many smaller groups rather than a few extremely large groups. This will not be an easy feat to accomplish, as the history of online social spaces has been a slow march toward increasing centralization, but it is worth considering how it might be achieved.

Note that some radical cyberlibertarians[2] have proposed a vision of the future of the social web with protocols in place of platforms, with fully distributed ownership of social data. In a sense, this is the extreme version of a user-driven approach, though it seems likely to be structured more as a network than as a set of groups. However, this is not the vision that I propose here. My vision for distributed social groups does not eliminate the role of the platforms, as I discuss below.

[2]Notably Mike Masnick https://web.archive.org/web/20200715213311/https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech

Each group involved in moderation engages in a process working toward cooperative responsibility

Much of the description of this aspect of my future vision can be found in Chapter 3, but I will briefly reiterate the major points. In defining cooperative responsibility, Helberger Pierson and Poell state that "platforms and users need to agree on the appropriate division of labor with regard to managing responsibility for their role in public space" [18, p. 2]. Though ultimately the division of responsibility should emerge from discussion between parties on each platform, my vision imagines a division of labor that follows roughly the following principles:

(1) Users should be responsible for being aware of the rules of each social space in which they plan to participate. There may be cases where violation of a rule is permissible or even necessary, but users must commit to taking their responsibility to the group and its rules seriously.

(2) Group moderators should be responsible for contributing to the group in a way that helps it achieve its purpose, which may evolve and change over time as the group grows or diversifies or external events impact its focus. Group moderators should also be responsible for the overall conduct of the group as it intersects with the rules of the platform; if a group repeatedly fails to meet the standards for behavior established for groups on the platform, then the moderators must be held responsible.

(3) Platform administrators should be responsible for general oversight of groups, but not oversight of individual users except in unusual cases (e.g., requests from law enforcement, cybersecurity threats, etc.). Platform administrators must ensure not only that groups are not antagonizing or disrupting other groups, but also that groups are meeting basic standards of decency in their internal operations. These standards must be cooperatively determined through the process outlined in Chapter 3, but platforms must commit to enforcing them both proactively and progressively. As many scholars have argued in depth (e.g., Gillespie [14, 15]), "neutrality" in content moderation is a harmful myth. Progressive content mod-

eration is a form of moderation that is context-aware, historically-informed, and committed to protecting the rights of marginalized users to participate on even footing with those who would (intentionally or otherwise) seek to silence them.

6.1.2 Pitfalls of community self-governance

As I have noted elsewhere, a community-driven approach to content moderation is not on its own a panacea for the problems with the modern content moderation ecosystem. There are a number of issues that must be addressed when considering potential future models for content moderation. The following are among the most important to consider:

The amplification of bad actors

Absent significant oversight, possibly in the manner prescribed for platforms above, all sorts of bad actors and problematic groups can take advantage of community self-moderation. White nationalist groups, anti-vaccine groups, incel groups, and groups that advocate for misogynistic violence can all exist in a purely community self-moderated ecosystem. These groups benefit as much as any other groups from the development of new and improved tools for community moderation. Moreover, these groups have historically shown a tendency to expand outward and attack and disrupt other groups (see, e.g., the examples from Reddit given in Chapter 3), and many have proven adept at taking advantage of platform affordances to do so. Contrary to the classic assertion that better speech will always triumph in a free marketplace of ideas, these groups understand that the speech that triumphs comes from the groups who are willing to spend the most resources to be the loudest. Thus, platforms that host self-governed communities will naturally amplify bad actors absent appropriate oversight.

Exploitation of free user labor

Community self-governance as a model for content moderation requires users to spend significant time and effort self-governing without receiving any material compensation for their efforts. If these communities exist on commercial platforms, as in my vision for the future they do, then these platforms are profiting from the free labor given by volunteer moderators who effectively improve the platform's product and increase other users' level of engagement with the platform. This is an ethically questionable labor arrangement, and it is one that is difficult to escape. Platforms are very unlikely to pay volunteer moderators; as I note later in this chapter, Facebook Groups alone has many millions of volunteer moderators, and paying them all would be both quite expensive and legally complicated. Alternative methods of compensation could be explored, such as on-site compensation (for example, one Reddit moderator interviewee suggested that it would be nice if Reddit moderators were given Reddit Gold for their efforts [34]) or access to additional premium features. Beyond financial issues, formalizing a relationship between platform administrators and volunteer moderators is organizationally complicated for reasons discussed in Chapter 3, namely that there are many understandable reasons that platforms do not always want to communicate openly with volunteer moderators. Given all the above, it is fair to question whether a content moderation model with such an ethically questionable labor arrangement at its core is worth exploring. I argue, however, that it is worth exploring because the alternatives are probably worse – despite the labor expected of them, moderators I interviewed almost universally said that they preferred to have the ability to moderate their own spaces and did not want to cede that authority to platforms (though they would have preferred slightly more support). In brief, the answer to a questionable arrangement of labor cannot be to take away agency from the (potentially) exploited party. I discuss this in more detail in the fifth guiding question below.

Reliance on Humans to Self-Govern

The final major danger in community self-governance is perhaps the most core – by definition, it relies on humans figuring out how to self-govern. Human intra- and inter-group social interactions often go poorly. Humans can be inefficient, lazy, jealous, rude, cruel, aggressive, incompetent, or clueless, and each of these qualities will make it more difficult for them to govern their social spaces. This is a classic social problem, and it means that many self-governed groups and communities will inevitably fail, fall apart, or descend into chaos. In considering this eventuality, we should keep in mind the broader social context – offline groups and communities have been failing since the dawn of human history for these same reasons, so the question here becomes "how much of a problem is it if groups and communities in online social spaces fail sometimes?" Or, in perhaps a more sophisticated train of thought – "what types of failures are acceptable, and can we design for online groups so as to minimize the unacceptable failures?" These questions are discussed in later sections of this chapter.

The above issues highlight several ways in which community self-governance is not inherently a utopian ideal, but rather one approach to content moderation with flaws and shortcomings. The following section considers how some of these shortcomings might be addressed through research.

6.2 Guiding Questions for Future Research

In this section I address research directions that can be pursued to make a self-moderated internet more of a reasonable possibility, building from the paths discussed above. I do not argue that self-moderation is the single best answer, nor that there is even a single answer that is best for every community, but rather that the self-moderation model has several significant advantages over the central moderation model and that the problems of the self-moderation model can be significantly ameliorated through further research and careful design.

(1) What are the factors that cause communities to be more or less amenable to a community self-moderation structure?

Various work has shown that larger communities rely more on automated tools and technical approaches to moderation, with, e.g., four times more moderation actions taken by bots than humans in large Twitch channels [33, pp. 157: 18–19], compared with a roughly even number taken by bots and humans in medium and small channels. Large communities can also rely extensively on user reports to guide moderators to potentially-problematic content [34, p. 14]. This parallels the struggles large platforms face in moderating huge volumes of content; they rely extensively on flags [8], automated tools, and contracted Commercial Content Moderators [31]. Despite the supposed ability of separate human review processes to be more context-sensitive, these processes can realistically only be more context-sensitive in a tiny fraction of cases because of the workload associated with understanding each case. Douek echoes this point with regard to the Facebook Oversight Board [11, pp. 5–6]. For these reasons and because of the sheer number of groups on the internet, context-sensitive self-moderation will require an enormous amount of human labor. However, though Facebook alone has hired tens of thousands of contractors to work full time moderating content on the site, this number is dwarfed by the number of users who already volunteer their time to keep self-moderated communities running; Facebook reports 200 million "meaningful" groups,[3] and each of these groups has at least one user who has volunteered to moderate.[4] If this volunteer force is to be able to take on a more significant role in the content moderation ecosystem, its relative strengths and weaknesses must be better understood.

A long history of literature supports the idea, at least implicitly, that the potential for nuanced, context-sensitive moderation depends significantly on the size of the community

[3] https://web.archive.org/web/20191208191626/https://singjupost.com/full-transcript-mark-zuckerberg-at-facebooks-f8-2018-developer-conference/?singlepage=1

[4] Note that the level of exploitation faced by these groups is clearly different; the former consistently face seriously problematic working conditions and often trauma [31].

in question; the MU* and BBS where scholars in the 1980s and 1990s observed public debates about rules and found pseudo-democratic processes typically had from hundreds of members up to a few thousand [23, 36, 35, 29]. On the other hand, the five communities from which moderators were interviewed for, e.g., Jhaver et al.'s 2019 analysis of Reddit's AutoModerator had subscriber counts ranging from 3.8 million to 17.4 million at the time of this writing.[5] However, having larger communities overall is not an inevitable result of the growth of the internet. For example, choosing arbitrary cutoffs, less than half a percent of live Twitch channels had more than 1000 current concurrent viewers at the time of writing, and approximately 90% of Twitch channels had 10 or fewer concurrent viewers.[6] Subreddits follow a similar power-law distribution, with less than half a percent of subreddits having more than 1000 subscribers and 75% having ten or fewer.[7] Though "subscribers" and "viewers" mean very different things, and the idea of a community looks very different on Reddit and Twitch, I can still be confident in concluding that the overwhelming majority of online communities are no bigger than the socially-developed communities studied by online community researchers in the 1990s.

Identifying a universal *maximum contextually-moderatable size* for online communities is impossible. Such a number would vary widely across types of platforms and types of communities within each platforms. Moderation in political discourse communities, for example, might become more difficult at a lower population threshold than in communities devoted to discussions of different types of trains. This threshold would also depend on available moderation tools, the number of moderators who have volunteered and how much time they can each commit, and various other factors. However, research that made at least preliminary attempts to model the concept of "moderatability" would be valuable. If community self-moderation is to become a more widely viable and, ideally, more widely adopted structure, one of two changes is required to the current structure of the internet: either (1) the

[5] https://web.archive.org/web/20200108024305/https://redditmetrics.com/top; subreddits from [19, p. 31:9]

[6] See also [33, p. 157:8] for a log-log plot of Twitch channel concurrent viewership as of mid-2018.

[7] https://redditmetrics.com/list-all-subreddits, accessed 7th January, 2020

largest online communities need to get smaller, need to be broken up, and/or a cap needs to be placed on community size; or (2), tools, systems, structures, and strategies need to be developed to increase the *maximum contextually-moderatable size* of communities. Of these two, the latter seems more realistic and is a place where researchers could contribute significantly, ideally with some form of modeling of this nebulous maximum size variable to track impact of certain types of interventions. Very basic questions that would contribute to the development of such a model have not yet been answered, e.g., "What is the impact on various moderation outcomes of adding one more moderator to a community's moderation team?" and "How much *work* does it take to moderate different types of communities?" Both of these questions could be operationalized in many different ways, but I am not aware of any quantitative research that has yet taken any approach to answering either.

In addition to size, a major factor impacting the viability of community-based approaches to moderation is platform structure. On many popular modern platforms, it would be complicated to directly impose a volunteer-based community moderation structure. Twitter, for example, has a core *network*-based structure, as opposed to a *community*-based structure. Twitter users don't join groups that are strictly bounded by any platform features; they follow individuals and have threaded conversations with multiple or many individuals. Despite these structural challenges, communities still do form on network-based platforms. Graham and Smith describe "Black Twitter", the large number of primarily African-American users who organize around the hashtag "#BlackTwitter" and various other related hashtags, as a *collective* with some aspects of a counterpublic [16]. Though Black Twitter does not have strictly-defined boundaries in terms of membership like, e.g., a closed Facebook Group, it can still be understood as a space or set of related spaces for discourse. Given this example, we could imagine one hypothetical case of community-based moderation on Twitter. What might it look like for Black Twitter to have volunteer moderators? Who would they be? How would they be chosen? What authority would they have, and what "spaces" would they have the power to shape? What might the negative consequences of such a structure

be? Similar questions can be asked about collectives or semi-cohesive communities on other networked platforms like the core Facebook network, Instagram, and perhaps even YouTube or TikTok. On the latter three platforms in particular, could the platforms design explicit structures to allow "collectives" of content creators who could formalize moderation processes and strategies within these new, composite communities?

(2) What are the processes of context-sensitivity in online community moderation, and how might they be better supported?

In writing about Facebook's struggle with how to moderate the famous "The Terror of War" photo depicting a young Vietnamese girl, naked and burned by napalm during the Vietnam War, Roberts argues that platforms are, by their nature, trapped into treating content as a commodity; their decisions regarding what content is to be permitted will thus be based on what content fits into processes of capital generation [30]. Such decision-making processes simply cannot be context sensitive in a company with a user base of nearly a third of the world's population. Douek sees the Facebook Oversight Board (Zuckerberg's "Supreme Court") as a partial answer to this problem, as, if properly designed, the Board will have the ability to examine local or even "hyper-local" context [11, pp. 35–36]. How, then, might the Board adjudicate an appeal of a takedown of a photo like "The Terror of War"? Presumably, they would consider the context in which the photo was posted and would issue an opinion describing whether they feel the image should have been taken down and why, and might perhaps go further and articulate a more general set of principles for when this image or this type of image should or should not be permitted. This opinion might in turn influence Facebook's overall policies, but it is very optimistic to believe that these proposed changes would impact Commercial Content Moderation processes in a way as nuanced as the Board originally intended; CCM workers simply do not have time to consider whether each instance of such an image meets criteria laid out in an Oversight Board opinion, and

the more nuanced the opinion, the less implementable it would be on a massive scale at the expected speed.

This scenario would play out much differently in a distributed moderation model where communities each made their own rules about content like "The Terror of War". In an offline context, it would be eminently reasonable for, e.g., a community orchestra to deem it inappropriate for members to wear T-shirts bearing this image to a performance, excepting highly specific and unusual circumstances. Similarly, at a young child's birthday party, parents might reasonably discourage guests from carrying around large reproductions of the photo. Though some people might find these rules unjust, few would argue that social groups should not be permitted to make their own rules and should only be beholden to centrally-determined standards for behavior. In a distributed online moderation model, similar processes would occur; moderators of a subreddit dedicated to liberal political discourse might find it very appropriate for a user to post an image of the photo, provided it was done in a respectful and thoughtful way. In contrast, moderators of a Facebook Group dedicated to sharing healthy eating tips might decide that this photo, while certainly important, wasn't appropriate for the group's context.

Though these examples are intentionally oversimplified, future research in community self-moderation could identify in more depth the factors that go into moderators' context-specific decisions about what to allow. Schoenebeck, Haimson, and Nakamura analyzed how users' identity attributes (e.g., race, class, political orientation) impact their attitudes toward different types of moderation actions taken by platforms [32]. Though these authors write with the goal of informing platforms' approaches, their work provides a strong starting point and model for similar work focusing on context-specificity in community self-moderation. For example, they explore users' attitudes not only toward bans and content removal but also toward approaches like apology, mediation, and payment, which are rooted in alternative theories of justice. Platforms that host communities that self-moderate have not traditionally been designed with these approaches in mind, but a deeper understanding of

how they currently play out could facilitate the design of spaces that support context-specific approaches to moderation.

A similar argument can be made in differentiating "conflict" from problematic behaviors. In work analyzing Wikipedia disputes, Billings and Watts argue that conflict is ever-present in online spaces, but that it "isn't all bad news". Conflict "can foster completely new perspectives on the activities and direction of the group, as the search for a resolution can produce new ways to conceptualize the issues at stake" [2, pp. 1447-1448]. Billings and Watts discuss the role of a subset of senior community members on Wikipedia, who they term "conciliators", who help guide editors who are in conflict to find their own solution to a dispute. They show how various social strategies used by conciliators, from restating the disputants' positions as they understand them to acknowledging social power differentials, can help scaffold conflict resolution. In an ideal system of distributed, community-based moderation, conflicts exist but are appropriately managed. Community moderators who are familiar with the contextual nuances of their communities' cultures are far better positioned to differentiate potentially-productive conflicts from problematic behaviors than platforms. They are also much more able to intervene with measures beyond what Schoenebeck, Haimson, and Nakamura call "traditional" moderation actions, i.e., bans and and content removal [32].

Conciliation is undeniably part of the role of moderators in many online spaces; the conflicts described above from LambdaMOO [9, 23] and MicroMUSE [35] both proceeded through mediation processes that involved participation by victims. More recently, Yu et al. found that moderators on MetaFilter use various approaches based on the concepts of *care* and *nurturing* to mediate conflicts and encourage constructive disagreement [40]. However, recent frameworks for community moderation have given less attention to these processes of conciliation. Seering et al.'s framework [34] describes moderation from a more functional perspective, generally taking moderators' authority as a given. Matias's framework [25] does note that moderators are in some sense accountable to members of their communities in

a civic sense, but does not discuss mediation or conciliation processes within communities. Future research that deepens understanding of the differential value of conflict and explores a variety of methods for managing conflict would help increase the viability of community-based approaches to online moderation.

(3) What is an "effective" moderator, and how does a person become one?

In their framework for work of volunteer moderators, Seering et al. present "Being and Becoming a Moderator" as one of their three main processes [34, pp. 8–12]. This process includes six sub-processes: Becoming a moderator, Role differentiation, Learning to be a moderator, Communication between moderators, Development of a moderation philosophy, and Relationship with site administrators. Seering et al. note that communication between moderators is a step that drives many of the other sub-processes; moderators learn and decide how to deal with different types of content by talking with other moderators and coming to consensus. Over time, these discussions lead to the development of moderation philosophies. However, formal processes for training new moderators were rare in their dataset. In most cases, moderators were expected to "learn by doing" or to work from an implicit understanding of the values and norms in the community. Despite this lack of formal structure, moderators were typically chosen *because* they exhibited a strong understanding of these values and norms; the most common reason that moderators were chosen was because they were standout members of the community.

Because of the variety of goals that moderators have, from maintaining a positive social environment [39] to generating useful and high quality political discussion [12] to taking care of a "fruit tree" [40] or nurturing a "garden" [34], "effectiveness" in moderation cannot be defined universally. However, future research could continue to hone a broader understanding of goals in community moderation and could classify how each of these goals fit into communities of different types. This work could predict, for a given community, what moderators'

goals would be, which could allow development of tools or recommendation systems tailored to specific contexts.

Individual moderators' processes for learning are also valuable to understand, but these processes are usually informal or nonexistent [34]. Certain sophisticated communities may have documentation or training processes like "trial periods", but this is not the norm. The creation of a body of knowledge about moderators' processes for learning, perhaps through ethnographic work, analysis of logs, or contextual inquiry, would allow for the development of tools to scaffold this learning. Though "learning by doing" is valuable in some situations, and it is important for moderators to have a chance to face the uncertainty that accompanies difficult decisions in order to develop a more sophisticated perspective, having access to support mechanisms could help new moderators acclimate to their roles more quickly or come to understand them more deeply.

(4) How can tools for moderation be developed that balance effectiveness with fairness and transparency?

Though, as discussed above, future research might be able to increase the maximum contextually-moderatable size of online communities, tools to automate or speed up the processes of moderation are likely to be necessary for the foreseeable future. For example, I can imagine a hypothetical system on Facebook Groups or Reddit that aggregates prior moderation decisions to predict whether a particular new piece of content is problematic. There are several ways that this might be operationalized: its training dataset could come from a general external source, as has been the case for much literature in the detection of hate speech, harassment, etc. [4, 7, 27], or it could come from aggregated data of past moderation decisions within the specific community in which it is deployed. The former approach can be influenced by significant bias from both the labelers and the source of the training data [3, 17], with offenses specific to the context of the host community unlikely to

be captured by a general dataset. In this sense, Chandrasekharan et al. [7, 5] have achieved some success by allowing moderators to build from the experiences of other communities, but questions of bias remain. The latter approach, drawing data from past decisions in the host community, requires a significant prior body of decisions and is unlikely to proactively predict changes in how content would be moderated due to external circumstances (e.g., a new type of problematic behavior appearing or a social change in the implicit meaning of a particular phrase or term); as machine learning models are trained on prior data, they are mathematically disadvantaged in making predictions on cases not contained in that data.

Tools can also granted variable agency to take different moderation actions, e.g., to remove content or ban users. Reddit's AutoModerator, for example, can be set to either flag or remove comments with certain words or phrases[8]. Flagging these comments allows moderators to decide for each case whether they should be permitted, while removing them relies on rules coded by moderators. Each approach has advantages and disadvantages; while allowing for human judgment in all cases likely increases attention to context and thus accuracy (for however normative accuracy can be defined), it can also increase moderators' workload significantly [19]. Algorithms that remove content without showing it to moderators first have another potentially-problematic effect: discussion between moderators about what content to permit and what to prohibit, both at a community's inception and as it grows, is a significant part of how communities evolve over time [34]. When a significant amount of content is automatically removed, moderators are less likely to have the chance to discuss its removal and thus it becomes more difficult for the relevant rules to evolve as contexts shift, and this problem is exacerbated as algorithms for detection become increasingly less transparent and can remove content without any clear explanation for why.

Chandrasekharan et al.'s Crossmod system [5] is a good example of work that thoughtfully considers ethical and technical tradeoffs; the system uses data from multiple communities on Reddit to train models for moderation, and allows moderators to customize various high-

[8] https://web.archive.org/web/20190728182154/https://www.reddit.com/wiki/automoderator

level aspects of the model to find a better fit for their community. Though the model itself remains somewhat opaque, the options for customization allow more supervision than a traditional externally-trained and blanket-applied model. The Crossmod system also refers flagged comments to moderators by default rather than directly removing them, which allows moderators to retain their agency. However, even this type of system presents challenges for handling questions of speech. For example, from a technical perspective, such a system is still naturally vulnerable to its mathematical limitations – a model trained on even very diverse data cannot make accurate predictions on content not in its training set, so novel problematic behaviors may pass unnoticed. Though these may be rare, the constantly-shifting nature of language use means that such a model will inevitably make inaccurate predictions in the first set of cases of new types of behavior. From a behavioral perspective, such a system could also encounter problems of both under-visibility of certain types of content and over-visibility of others; moderators could become dependent on algorithmic detection methods, spending less time browsing and flagging content on their own and relying more on the algorithms to identify problematic behavior. Thus, problematic behaviors not easily detected by algorithms might be more likely to go unmoderated. On the other hand, it is possible that the fact that an algorithm had flagged a piece of content might make moderators more likely to see it as deserving removal, where if they had seen it in its natural context without any particular flag, they might not have seen it as problematic. The algorithm's flag might serve as a pseudo-"anchor" in the psychological sense. Chandrasekharan et al. report that they brought comments flagged by Crossmod that hadn't previously been removed to moderators' attention, and moderators agreed that the vast majority of these did merit removal, but this review does not appear to have had a "control" condition; moderators were aware that all of these comments had already been flagged [5, pp. 174:23–24].

None of the above should be interpreted to mean that algorithmic approaches to content moderation should not be pursued. Though there are serious questions about the ethical and technical tradeoffs in these approaches, they are a necessary component of the modera-

tion ecosystem. Further research in this area should focus in depth on understanding these tradeoffs and developing approaches that attempt to mitigate the associated harms.

(5) How can the relationship between volunteer community moderators and platform administrators be made more productive?

The practice of volunteer community moderation is, from a perspective of labor, a questionable one. If, per Gillespie, moderation is "*the* commodity that platforms offer" [15, p. 13], the business model of platforms like Facebook and Reddit depends significantly on the labor of volunteers who receive none of the profits that result. This business model, when combined with the idea of moderator "burnout", the process by which moderators become exhausted, drained, or overwhelmed by their work and possibly quit as a result [10], makes reliance on users' labor seem somewhat exploitative. In a broader discussion on the concept of free labor online, Terranova argues that the internet is "the latest capitalist machination against labor", and makes broad connections between labor, politics, and culture [37, p. 54]. However, it is important to consider moderators' voices before imposing the label of exploitation on their working conditions. While some moderators wish the platform recognized their efforts in at least a token way, few feel at all trapped or pressured to be in their position [34, pp. 20]. These moderators typically volunteer because their community is meaningful to them or because they feel like they could gain something from being a moderator [39]; like students volunteering to organize drama clubs, parents volunteering to be Girl Scout troop leaders, or community members leading Bible study groups, volunteering to lead is a core part of human socialization, both online and offline. Though the profit structures are certainly different for, e.g., groups run on Facebook vs groups run locally in a geographically-bounded community, many offline volunteer-run communities are also in some ways dependent on businesses or have businesses built around them. Broadly, the manner in which moderators' free labor contributes to platforms' profit is questionable and potentially exploitative, but this fact

alone does not justify the argument that users' ability to self-moderate should be taken from them and given to platforms.

In considering the organizational structures of moderation and comparing them with offline volunteer organizations and Groups on Facebook or Reddit, one important structural difference to note is the relationship between volunteer moderators and platform administrators. Ostensibly, both groups' interests in terms of moderation are fairly-well aligned in a general sense; both groups want a platform where people can socialize and interact in reasonable, positive ways. Both groups, at least by majority of members, want a platform free from behaviors like stalking, extreme harassment, covert political influence operations, and targeted cyber-attacks. However, despite both being involved in moderation processes with similar goals, the two groups do not have effective structures set up for communication. Regular users [38] and even moderators can have content removed in their communities by platforms and have no idea why, and appeals processes rarely yield clear explanations. Platforms also regularly conceal their broader actions and motives from volunteer moderators, as in the case of Reddit's regular banning of accounts they suspect to be associated with government-sponsored influence operations. It is understandable that Reddit would not share its methods or progress in doing this, as doing so could make it much easier for bad actors to evade detection; the wide variety of calls for transparency in platform moderation [15, 21] acknowledge the complexity of sharing information, but none provide a clear pathway for managing an organizational structure that includes both volunteer moderators and platforms and balancing self-moderation with exploitation.

This inherent clash between platforms' and volunteers' motives and processes is perhaps the most difficult challenge in propagating a community-based moderation model. Many volunteer-run communities studied in the 1980s and 1990s did not have this problem, as they were independently-formed and not beholden to any particular company, while the vast majority of modern online socialization happens on platforms run by businesses. These platforms are unlikely to formally recognize volunteer moderators as employees, contractors,

or even laborers in any formal sense, as to do so would invite broad legal and structural challenges to their processes. One direction worth exploring, however, is a more clear separation of responsibilities between platforms and volunteer moderators, which aligns well with Helberger, Pierson, and Poell's concept of "cooperative responsibility" [18, p. 2]. On Twitch and to some extent Reddit, the platforms expect volunteers to do day-to-day moderation of individual pieces of content, intervening primarily only when a full community is seen as problematic or when problematic behaviors from one community spill over and impact other communities [6, 24]. Reasonable critiques can be made about when these platforms have (and have not) decided to intervene, e.g., Massanari's assertion that there is a "deep reluctance on the part of the [Reddit] administrators to alienate any part of their audience, no matter how problematic" [24, p. 340], but this divided-labor model does grant more autonomy to users to decide for themselves what is appropriate in their spaces and also more clearly delineates the domains of platforms and users in moderation.

(6) Can more democratic mechanisms, e.g., referendums, appeals processes, etc., be effectively (re) integrated into online communities? If so, how?

In his article, "Is Democratic Leadership Possible?" Beerbohm reconsiders whether the concept of leadership is incompatible with our conception of democracy [1]. If a leadership is when "an agent gets a collection of agents to do or believe something without coercion" [p. 639], then leadership is antithetical to a view of democracy where leaders are directly responsive to preferences of constituents. Moderators in online communities are certainly not democratic leaders within this definition; the literature above shows that they engage in a variety of processes that are intended to shape the trajectories of communities, and even if they were democratically elected it seems unlikely that they would act purely as representatives of the preferences of their "constituents". Beerbohm provides an alternative theory of leadership, the Commitment Theory, which argues that "Democratic leadership's

success condition is the recruitment of citizens as genuine partners in shared political activity" [p. 639]. Leaders can encourage or persuade constituents to adopt a particular commitment to address an issue in a certain way, so long as this is done without deception, and they can then coordinate political actions toward achieving this commitment [p.641–644].

Though Beerbohm focuses on systems that fall more traditionally within the fields of political science and government, this definition of democratic leadership applies well to online communities and offers a path forward for design. Moderators can consider a particular situation, identify what they believe to be the best approach, and attempt to create a joint commitment with community members toward addressing the situation in that way, though they should enter conversation with community members with a genuine intent to listen to community members and be receptive to their concerns and even to change the proposed approach as a result. Though democratic leadership does, realistically speaking, require elections of some sort, the type of leadership described in the Commitment Theory can be incorporated into community governance even in the intermediate step prior to establishment of elections. Moderators who attempt to work with community members as partners are certainly closer to the ideal of democratic leadership than those who make decisions without any input from their communities.

The examples of the emergence of democratic principles on LambdaMOO [23], MicroMUSE [35], and Lucasfilm's Habitat [26] show that, given the proper conditions, more democratic moderation of online communities is possible, but major social platforms have not been designed to support these processes.[9] Though creating structures for democratically-driven moderation is a challenging problem, it is one that is possible to engage empirically. There is no shortage of online communities with which collaborative studies can be run, and inspiration for feature designs can be taken both from the examples of communities in early literature and, to some extent, from peer production spaces like Wikis and FOSS development communities. Frey, Krafft, and Keegan [13] draw on Ostrom's work on scaffolding

[9]Slashdot, as reported by Lampe et al. [22], allowed for meta-moderation, where users rated the fairness of other users' moderation decisions, but these features have not been widely adopted.

democratic participation within communities (e.g., [28]) to argue for an increased focus on development of a body of research based on real-world studies that could lead to a "science of digital institution design". This science could help designers understand how users might fill more substantive roles in platform moderation. Though this line of research, along with the many other possible complimentary approaches, could require a breadth of skills not usually present in single domains, ranging from qualitative inquiry to design and development of systems and running of experiments, it is an excellent fit for interdisciplinary research domains. Though Frey, Krafft, and Keegan's "digital institution design" is mostly focused on structures for user participation, it is worth considering the potential for user ownership of interface design as a whole, from buttons to background colors to modes of communication. Could platforms like Facebook "design" their social spaces so they could truly be designed by users, whether via full individual customization or collaborative construction? The history of online communities is full of examples where social spaces were collaboratively constructed in this way; prior to its current iteration, where flexibility has been reduced, even Reddit allowed moderators significant leeway to customize subreddit layouts and modify a small set of features.

It is unclear whether Beerbohm's definition of democratic leadership is possible to achieve in online spaces at any scale [1]. Even in the spaces discussed in early research, the implementation of democratic systems only partially checked the power of moderators; the resulting moderation structures were still closer to oligarchies than democracies or even democratic republics. Systems of democratic leadership may not be necessary for the smooth functioning of online social spaces; many offline social structures are run by people who are not democratically elected. However, democratic principles, even if not strictly necessary for "effective" moderation, are worth exploring both as a complex design challenge and as an important social question that can illuminate new ways to grant users power and agency in self-moderation.

6.3 Concluding thoughts

In their 2018 article on governance of the broad space of online platforms, Helberger, Pierson, and Poell argue that platforms have not lived up to their promise:

> *[Modern] platforms typically appear to facilitate public activity with very little aid from public institutions. As such, they are celebrated as instruments of what has become known as "participatory culture" and the "sharing" or "collaborative" economy [...] Online platforms hold the promise of empowering individuals to effectively take up their role as producers of public goods and services, as well as to act as autonomous and responsible citizens. However, in practice, online platforms have, to date, not fulfilled this promise. Instead, in many cases they appear to be further intensifying the pressure of the market on important public values, such as transparency and non-discrimination in service delivery, civility of public communication, and diversity of media content.* [18, p. 1]

In the conclusion to her 2018 article discussing Facebook's response to "The Terror of War", Roberts takes an even more critical stance on the future of platform-based social media:

> *Perhaps the problem is so deeply structural that spaces like Facebook and other UGC-reliant advertising platforms, by virtue of their own ecosystem made up of architecture and functionality, economy and policy, ultimately suffer from an inability to convey any real depth of meaning at all. Under these circumstances, the utility of platforms, governed by profit motive and operating under a logic of opacity, to the end of greater ideals or challenge to status quo is seriously in doubt.* [30]

Though I acknowledge the important role these platforms have played in a variety of movements for social change, these authors' central points are fair; the moderation of public

discourse under a centralized, profit-driven system deserves heavy scrutiny. Whether these models can truly achieve the values of transparency and accountability that have been articulated as goals for their future remains to be seen. As I have argued throughout this work, though many recent, well-received pieces of research have focused exclusively on platform-driven moderation without considering alternative moderation models, the platform-driven model should not be treated as an inevitable outcome of the current trends; there are many forms of moderation that do not fall under this centralized model, and many of these systems already exist and have seen significant success. The modern world of social media, in its current, centrally-moderated form, is certainly not the internet-facilitated Utopia that we were promised, but an open-minded reconsideration of the core structures of the modern internet may lead us toward viable alternatives.

Bibliography

[1] Eric Beerbohm. Is democratic leadership possible? *American Political Science Review*, 109(4):639–652, 2015.

[2] Matt Billings and Leon A. Watts. Understanding dispute resolution online: Using text to reflect personal and substantive issues in conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1447–1456, New York, NY, USA, 2010. ACM.

[3] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[4] 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 & Internet*, 7(2):223–242, 2015.

[5] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[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. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):31:1–31:22, December 2017.

[7] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3175–3187, New York, NY, USA, 2017. ACM.

[8] 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.

[9] 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*, December 23:36–42, 1993.

[10] Bryan Dosono and Bryan Semaan. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[11] Evelyn Douek. Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility. *N.C. J.L. & Tech*, 21:1–78, 2019.

[12] Dmitry Epstein and Gilly Leshed. The magic sauce: Practices of facilitation in online policy deliberation. *Journal of Public Deliberation*, 12(1), 2016.

[13] Seth Frey, P. M. Krafft, and Brian C. Keegan. "this place does what it was built for": Designing digital institutions for participatory change. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[14] Tarleton Gillespie. The politics of 'platforms'. *New Media & Society*, 12(3):347–364, 2010.

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

[16] Roderick Graham and Shawn Smith. The Content of Our #Characters: Black Twitter as Counterpublic. *Sociology of Race and Ethnicity*, 2(4):433–449, 2016.

[17] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, pages 2–12, New York, NY, USA, 2018. ACM.

[18] Natali Helberger, Jo Pierson, and Thomas Poell. Governing online platforms: From contested to cooperative responsibility. *Information Society*, 2018.

[19] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5):31:1–31:35, July 2019.

[20] David R. Johnson and David Post. The new 'civic virtue' of the internet. *First Monday*, 3(1), 1998.

[21] David Kaye. *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports, New York, NY, USA, 2019.

[22] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317–326, 2014.

[23] Richard MacKinnon. Virtual rape. *Journal of Computer-Mediated Communication*, 2(4):1–2, 1997.

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

[25] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 5(2), 2019.

[26] Chip Morningstar and F Randall Farmer. The Lessons of Lucasfilm's Habitat. In Michael Benedikt, editor, *Cyberspace: First Steps*, pages 273–301. MIT Press, Cambridge, MA, USA, 1991.

[27] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[28] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK, 1990.

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

[30] Sarah Roberts. Digital detritus: 'error' and the logic of opacity in social media content moderation. *First Monday*, 23(3), 2018.

[31] Sarah T Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, CT, USA, 2019.

[32] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. Drawing from justice theories to support targets of online harassment. *New Media & Society*, page 1461444820913122, 2020.

[33] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. The social roles of bots: Evaluating impact of bots on discussions in online communities. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):157:1–157:29, November 2018.

[34] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media & Society*, 21(7):1417–1443, 2019.

[35] 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, New York, NY, USA, 1st edition, 1999.

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

[37] Tiziana Terranova. Free labor: Producing culture for the digital economy. *Social text*, 18(2):33–58, 2000.

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

[39] Donghee Yvette Wohn. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 160:1–160:13, New York, NY, USA, 2019. ACM.

[40] Bingjie Yu, Katta Spiel, Joseph Seering, and Leon Watts. "Taking Care of a Fruit Tree": Nurturing as a Layer of Concern in Online Community Moderation. In *CHI '20 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '20, pages 253:1–9, New York, NY, USA, 2020. ACM.