Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation

JOSEPH SEERING, Carnegie Mellon University, USA

Research in online content moderation has a long history of exploring different forms that moderation can take, including both user-driven moderation models on community-based platforms like Wikipedia, Facebook Groups, and Reddit, and centralized corporate moderation models on platforms like Twitter and Instagram. In this work I review different approaches to moderation research with the goal of providing a roadmap for researchers studying community self-moderation. I contrast community-based moderation research with platforms and policies-focused moderation research, and argue that the former has an important role to play in shaping discussions about the future of online moderation. I provide six guiding questions for future research that, if answered, can support the development of a form of user-driven moderation that is widely implementable across a variety of social spaces online, offering an alternative to the corporate moderation models that dominate public debate and discussion.

CCS Concepts: • **Human-centered computing → HCI theory, concepts, and models**.

Additional Key Words and Phrases: Online communities; Moderation; Governance; Social Networks; Platforms; Harassment; Hate speech

ACM Reference Format:
Joseph Seering. 2020. Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 107 (October 2020), 28 pages. https://doi.org/10.1145/3415178

1 INTRODUCTION

In May of 1978, the “CommuniTree #1” online Bulletin Board System (BBS) launched in the San Francisco Bay area [95, pp. 88–92], [96]. 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 [96], 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

Author’s address: Joseph Seering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, USA, jseering@andrew.cmu.edu.

2573-0142/2020/10-ART107
https://doi.org/10.1145/3415178

or participants in a novel discourse but as agents of a new kind of social experiment" [95, p. 90][1]. In 1982, Apple entered into an 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" [96], 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,[2] or adopt defensive behaviors [29, 102]. A 2017 Pew study found that 4 in 10 Americans had personally experienced online harassment [79]. Major platforms have closed their comment sections or forums because of an inability to maintain positive and productive conversations, including NPR,[3] Popular Science,[4] and IMDB.[5] 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.

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,[6] 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. 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

[1] Reflecting the "Digital Utopianism" [100] 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." [36, p. iii]

[2] 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, and https://web.archive.org/web/20190407195949/https://www.polygon.com/2016/7/19/12222842/ghostbusters-leslie-jones-twitter-abuse on Twitter.

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

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

[5] https://web.archive.org/web/20190410195135/https://www.theverge.com/2017/2/3/14501390/imdb-closing-user-forums-comments

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

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 paper is to provide a clear outline of the state of community moderation research in the context of other domains and to suggest a path for how researchers in this field might come to articulate a clear model of a functional internet where community self-moderation plays a greater role.

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?' [22, 97] and 'How might laws be written to shift platforms' content moderation processes in more productive directions?' [15]. 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?" [68, 105] and "How do these moderators make rules?" [90]. Note that these perspectives are not mutually exclusive, and a valuable goal for future work would be to integrate the two.[7] 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.[8] 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 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.

2.1 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.

2.1.1 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 [40]. 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*

[7] See, e.g., guiding question number five.

[8] 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.

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"' [39], 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 [58]. 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" [16, 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.[9] 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, 24]. 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 moderation; Danielle Keats Citron's *Hate Crimes in Cyberspace* discusses the role of Section 230 in protecting platforms that host "revenge porn" [14]. 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 [30, pp. 1282–1291].[10] 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 [15, 16] [31, 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 [15, 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

[9] 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

[10] 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.

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 [22] 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" [22, 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 makes a number of proposals for improving content moderation [54, 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" [54, 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.

2.1.2 Structural and Functional Perspectives: In his 2015 work, "The Virtues of Moderation", Grimmelmann [42] 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 [18]. 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. [7] 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 [18, 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 [78]. 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.[11] 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 [103]. Suzor et al. proposed specific ways in which increased transparency could help educate users and establish a sense of trust in these processes [97]. Achieving transparency is a significant challenge; increased transparency would allow for more public and perhaps democratic 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" [84]. 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.[12] 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 [85, p. 50].

2.1.3 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 [12], 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.

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 [33], a phenomenon further explored in research by Jhaver et al. [51], 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 [38] proposed

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

[12] Gillespie notes that, in its early days, Facebook relied on Harvard students to volunteer their time as moderators [40, p. 118].

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 [8], Instagram [64], and online news comments [72].[13] 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. [72] focus on hate speech, but do not provide a specific definition by which raters applied this label. Burnap and Williams [8, 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. [64, 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 data they were trained on [43]. 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 [6]. 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 [85] 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,[14] 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 The Communities Perspective: Users' Intra-group Moderation

In studying the processes of moderation in online communities, scholars including Viégas et al. [101], Forte, Larco, and Bruckman [28], and Kollock and Smith [59] have drawn on a framework proposed by Ostrom for "Design Principles of Long-Surviving, Self-Organized Resource Regimes" [75], [76, 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

[13] 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.

[14] 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.

Lanka, Nova Scotia, and the United States [74]. Ostrom's work explored governance systems that she explicitly framed as alternatives to firm and state-driven approaches to governance [74, pp. 8–20, 40–45], [77, pp. 10–13]. I argue, as did Ostrom, that there is "an alternative way" of addressing problems of community governance [74, 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., [48]) and peaking in volume in the mid-to-late 1990s. This work focused primarily on spaces like Usenet and MUDs[15] (e.g., [20, 47, 65, 81, 92, 94]), but also explored spaces like Electronic Bulletin Board Systems [82, pp. 131–144], [95], the Whole Earth 'Lectronic Link (WELL) [82, pp. 18–64], Lucasfilm's Habitat [69], and numerous other smaller spaces. These platforms were largely 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, 4] including Wikipedia (e.g., [5, 28, 35, 57, 91]) and free and open-source software (FOSS) communities (e.g., [55, 73]). A smaller body of work in this era focused on "distributed" or "crowdsourced" moderation [62, 63], which would provide a foundation for later analysis of platforms like Reddit.[16] 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 [25, 71], the McGill Online Design Studio [25], and British government-run online discussion fora [106].

These platforms have fundamentally different characteristics than the major social media platforms that Gillespie [40], Klonick [58], 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.

2.2.1 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 [61] 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 in spaces that are well-moderated [104], and thoughtful moderation can increase the quality of users' contributions [17] and help steer communities through periods of turbulence

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

[16] See Kou et al.'s work [60] on "crowdsourced" moderation in League of Legends for another more recent example.

[56, 80]. Though platform administrators[17] 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 [90, p. 11].

Though the practices of these moderators are diverse and complex, Seering et al. [90] 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. [90] and Squirrel's work on Reddit moderation [93] has explored the second, but I discuss the second later in this paper as an area that merits further research.

2.2.2 The everyday labor of volunteer moderators: In his 2016 work "The Civic Labor of Online Moderators" [68], Matias builds on literature from Gillespie [39], Shaw [91], Grimmelman [42], 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" [68, 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" [40, 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" [105, 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" [23, 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 without seeing at least an occasional ban or content removal. On some platforms, community members participate directly in

[17] 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 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 [18], 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 [23, pp. 4: 10–12], [44].

Though a significant portion of moderators' work can require tools or features, Seering et al. [90] 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 [9, pp. 167–169]. Other work has identified similar strategies across a wide variety of platforms, including both social spaces [90] and spaces for political discourse [23, 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) [90]. 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 [50]. 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 [90, p. 17]. These social strategies are not new, and are implicit in much prior work on moderation (e.g., [92, 94]), 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 paper. In her studies of "MicroMUSE", performed in the early-to-mid 1990s, Smith [92, 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.[18] "Followers-only mode" on Twitch requires users to have been present for a certain amount of time before they are allowed to post [90], and Automoderator settings on Reddit can also prohibit new users from posting in certain communities [49]. Wikipedia maintains an "Adopt-a-user" mentoring program that pairs new users with more experienced Wikipedians [70].

2.2.3 Guiding and shaping a community: The second theme that Matias notes is "Moderation as Civic Participation" [68, 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" [90, 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 [90]. Wohn focuses in depth on these types of civic motivations [105, 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* [94, 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 [90, p. 16–19]. Building from work on Reddit, Squirrel [93] 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 [90]. 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.[19] Accordingly, the final major theme that Matias identifies is "Moderation as Oligarchy" [68, 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 [65] described a widely-cited incident originally reported by Dibbel [19] of community "justice" in a

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

[19] As Forte, Larco, and Bruckmam note, Wikipedia's decision-making processes could in some senses be called Democratic [28], but as the recruitment of new editors has slowed, Wikipedia has become decidedly more oligarchic [45, 91].

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 [19, 65], 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 [92, 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 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 [92, 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" [69, 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 [69, 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 [40] 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 [53] 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 [50], 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 [10, 37].

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 [90]. 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 [67, 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 [90]. 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.

2.2.4 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 [11]. 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 [90, 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. [49] and Seering et al., it is important to develop tools that "support, rather than supplant, the judgment of users" [90, p. 3], and, per Geiger [34], 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

[66] 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. [7] 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 [87]. Other recent work has employed "empathy nudges" to attempt to encourage bystander intervention, with mixed results [98]. 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 [89]. 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.

3 DISCUSSION AND QUESTIONS 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.

> *We 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.* [52]

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) 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 [54] 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. For example, the Archive of Our Own (AO3) fanfiction writing community, which hosts over two million users and more than five million works of fanfiction, was created in response to disagreement with decisions made by Livejournal and FanLib, two for-profit sites that had previously been major hubs for fanfiction. AO3 was designed and coded primarily by fan volunteers, most of whom were women, based on feedback from their community. It incorporates various custom-designed features for writing and filtering and tagging, with special attention paid to accessibility and inclusivity [27, pp. 2574–2579]. Though fanfiction operates in a complex legal and ethical space, AO3 developers have been able to embed community norms directly into the platform's code. For example, per the fanfiction community's norm of giving credit to work that influenced one's writing, AO3 has a feature to add a "citation" directly into an article's metadata to credit work that inspired it [26, pp. 241:8].

3.1 Guiding Questions for Future Work

In this section I address research directions that can be pursued to make a self-moderated internet more of a reasonable possibility. 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 [88, 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 [90, p. 14]. This parallels the struggles large platforms face in moderating huge volumes of content; they rely extensively on flags [18], automated tools, and contracted Commercial Content Moderators [85]. 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 [22, 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,[20] and each of these groups has at least one user who has volunteered to moderate.[21] 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 in question; the MUDs 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 [65, 81, 92, 95]. 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.[22] 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.[23] 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.[24] Though "subscribers" and "viewers" mean very different things, and the idea of a community looks very different on Reddit and Twitch, it is clear that the overwhelming majority of modern online communities are no bigger than the 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 and various other factors. However, research that made at least preliminary attempts to model the concept of "moderatability" would be valuable. 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?"

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 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 [41]. 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

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

[21] Note that the level of exploitation faced by these groups is clearly different; the former consistently face seriously problematic working conditions and often trauma [85].

[22] https://web.archive.org/web/20200108024305/https://redditmetrics.com/top; subreddits from [49, p. 31:9]

[23] See also [88, p. 157:8] for a log-log plot of Twitch channel concurrent viewership as of mid-2018.

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

would they be chosen? What authority would they have? 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 [83]. 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 [22, 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 [86]. 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” [5, 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. 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 [86].

Conciliation is undeniably part of the role of moderators in many online spaces. However, recent frameworks for community moderation have given less attention to these processes of conciliation. Seering et al.’s framework [90] describes moderation from a more functional perspective, generally taking moderators’ authority as a given. Matias’s framework [68] 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 [90, pp. 8–12]. This process includes six subprocesses connected largely by communication between moderators; 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 [105] to generating useful and high quality political discussion [23] to taking care of a “fruit tree” [107] or nurturing a “garden” [90], “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 [90]. 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, we could 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. [8, 13, 72], 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 [6, 43], with offenses specific to the context of the host community unlikely to be captured by a general dataset. In this sense, Chandrasekharan et al. [11, 13] 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 be 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.[25] 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 [49]. 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 [90]. 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 [11] 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-level aspects of the

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

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 means that such a model will make less accurate 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 [11, 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 moderation 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" [40, 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 [21], 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 [99, 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 [90, 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 [105]; 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 [103] 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 [40, 54] 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" [46, 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 [12, 67]. 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" [67, 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 [2]. 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 be receptive to community members' concerns. 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 [65], MicroMUSE [92], and Lucasfilm's Habitat [69] 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.[26] 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 [32] draw on Ostrom's work on scaffolding democratic participation within communities (e.g., [74]) 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?

It is unclear whether Beerbohm's definition of democratic leadership is possible to achieve in online spaces at any scale [2]. 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;

[26] Slashdot, as reported by Lampe et al. [63], allowed for meta-moderation, where users rated the fairness of other users' moderation decisions, but these features have not been widely adopted.

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.

4 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.* [46, 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.* [83]

Though how Facebook and other platforms impact public discourse is far from a settled question, 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 paper, 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.

ACKNOWLEDGMENTS

This work was possible because of support and feedback from numerous people, including Geoff Kaufman, Greg MacDonough, April Sperry, Warren Seering, Adam Seering, and Nancy Cable. I would also like to thank the reviewers for their thorough comments.

REFERENCES

[1] Jan Philipp Albrecht. 2016. How the GDPR Will Change the World. *European Data Protection Law Review* 2, 3 (2016), 3. https://doi.org/10.21552/EDPL/2016/3/4

[2] Eric Beerbohm. 2015. Is Democratic Leadership Possible? *American Political Science Review* 109, 4 (2015), 639–652. https://doi.org/10.1017/S0003055415000398

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

[4] Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2015. Peer Production: A Form of Collective Intelligence. In *Handbook of Collective Intelligence*, Thomas Malone and Michael Bernstein (Eds.). MIT Press, Cambridge, MA, USA, 175–204.

[5] Matt Billings and Leon A. Watts. 2010. 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)*. Association for Computing Machinery, New York, NY, USA, 1447–1456. https://doi.org/10.1145/1753326.1753542

[6] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Social Informatics. SocInfo 2017. Lecture Notes in Computer Science, vol 10540*, G. Ciampaglia, A. Mashhadi, and T. Yasseri (Eds.). Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-67256-4_32

[7] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134659

[8] Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet* 7, 2 (2015), 223–242. https://doi.org/10.1002/poi3.85

[9] Jie Cai and Donghee Yvette Wohn. 2019. 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)*. Association for Computing Machinery, New York, NY, USA, 166–170. https://doi.org/10.1145/3311957.3359478

[10] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 174 (Nov. 2019), 30 pages. https://doi.org/10.1145/3359276

[12] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666

[13] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. 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)*. ACM, New York, NY, USA, 3175–3187. https://doi.org/10.1145/3025453.3026018

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

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

[16] Danielle Keats Citron and Benjamin Wittes. 2018. The Problem Isn't Just Backpage: Revising Section 230 Immunity. *Georgetown Law Technology Review (2018)* 2 (23 July 2018), 453–473. Issue 2.

[17] Dan Cosley, Dan Frankowski, Sara Kiesler, Loren Terveen, and John Riedl. 2005. How Oversight Improves Member-maintained Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/1054972.1054975

[18] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163

[19] Julian Dibbell. 1993. 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 (1993), 36–42. https://www.villagevoice.com/2005/10/18/a-rape-in-cyberspace/

[20] Judith Donath. 1999. Identity and Deception in the Virtual Community. In *Communities in Cyberspace* (1st ed.), Marc A Smith and Peter Kollock (Eds.). Routledge, London, UK, 27–58. https://doi.org/10.1519/JSC.0b013e3181e4f7a9

[21] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 142, 13 pages. https://doi.org/10.1145/3290605.3300372

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

[23] Dmitry Epstein and Gilly Leshed. 2016. The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. *Journal of Public Deliberation* 12, 1, Article 4 (2016), 29 pages.

[24] European Commission. 2016. Regulation (EU) 2016/679 (General Data Protection Regulation). OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf

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

[26] Casey Fiesler and Amy S. Bruckman. 2019. Creativity, Copyright, and Close-Knit Communities: A Case Study of Social Norm Formation and Enforcement. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 241 (Dec. 2019), 24 pages. https://doi.org/10.1145/3361122

[27] Casey Fiesler, Shannon Morrison, and Amy S. Bruckman. 2016. An Archive of Their Own: A Case Study of Feminist HCI and Values in Design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2574–2585. https://doi.org/10.1145/2858036.2858409

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

[29] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19, 8 (2017), 1290–1307. https://doi.org/10.1177/1461444816635778

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

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

[32] Seth Frey, P. M. Krafft, and Brian C. Keegan. 2019. "This Place Does What It Was Built For": Designing Digital Institutions for Participatory Change. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article Article 32 (Nov. 2019), 31 pages. https://doi.org/10.1145/3359134

[33] R. Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803. https://doi.org/10.1080/1369118X.2016.1153700

[34] R Stuart Geiger. 2017. Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society* 4, 2 (2017), 1–14. https://doi.org/10.1177/2053951717730735

[35] R. Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 117–126. https://doi.org/10.1145/1718918.1718941

[36] Dean Gengle. 1981. *Communitree* (first ed.). The CommuniTree Group, San Francisco, CA, USA.

[37] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511. https://doi.org/10.1177/1461444818776611

[38] Arpita Ghosh, Satyen Kale, and Preston McAfee. 2011. Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC '11)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/1993574.1993599

[39] Tarleton Gillespie. 2010. The politics of 'platforms'. *New Media & Society* 12, 3 (2010), 347–364. https://doi.org/10.1177/1461444809342738

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

[41] Roderick Graham and Shawn Smith. 2016. The Content of Our #Characters: Black Twitter as Counterpublic. *Sociology of Race and Ethnicity* 2, 4 (2016), 433–449. https://doi.org/10.1177/2332649216639067

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

[43] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All You Need is "Love": Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec '18)*. Association for Computing Machinery, New York, NY, USA, 2–12. https://doi.org/10.1145/3270101.3270103

[44] David Gurzick, Kevin F. White, Wayne G. Lutters, and Lee Boot. 2009. 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)*. Association for Computing Machinery, New York, NY, USA, 361–370. https://doi.org/10.1145/1531674.1531727

[45] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5 (2013), 664–688. https://doi.org/10.1177/0002764212469365

[46] Natali Helberger, Jo Pierson, and Thomas Poell. 2018. Governing online platforms: From contested to cooperative responsibility. *Information Society* 34, 1 (2018), 1–14. https://doi.org/10.1080/01972243.2017.1391913

[47] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society* 18, 5 (2002), 371–384. https://doi.org/10.1080/01972240290108186

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

[49] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243

[50] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (Nov. 2019), 27 pages. https://doi.org/10.1145/3359252

[51] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593

[52] David R. Johnson and David Post. 1998. The New 'Civic Virtue' of the Internet. *First Monday* 3, 1 (1998), 18. https://doi.org/10.5210/fm.v3i1.570

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

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

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

[56] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 1152–1156. https://doi.org/10.1145/2858036.2858356

[57] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07).* ACM, New York, NY, USA, 453–462. https://doi.org/10.1145/1240624.1240698

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

[59] Peter Kollock and Marc Smith. 1996. Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. In *Computer-mediated Communication: Linguistic, Social, and Cross-cultural Perspectives*, Susan Herring (Ed.). John Benjamins Publishing, Amsterdam, Netherlands, 109–128.

[60] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior Through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 62 (Dec. 2017), 17 pages. https://doi.org/10.1145/3134697

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

[62] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04).* ACM, New York, NY, USA, 543–550. https://doi.org/10.1145/985692.985761

[63] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (2014), 317–326. https://doi.org/10.1016/j.giq.2013.11.005

[64] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. 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)*, Vol. 91. AAAI, Menlo Park, CA, USA, 181–190.

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

[66] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. 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). ACM, New York, NY, USA, Article 586, 13 pages. https://doi.org/10.1145/3173574.3174160

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

[68] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (2019), 12. https://doi.org/10.1177/2056305119836778

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

[70] David R. Musicant, Yuqing Ren, James A. Johnson, and John Riedl. 2011. Mentoring in Wikipedia: A Clash of Cultures. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. ACM, New York, NY, USA, 173–182. https://doi.org/10.1145/2038558.2038586

[71] Chaebong Nam. 2019. Behind the interface: Human moderation for deliberative engagement in an eRulemaking discussion. *Government Information Quarterly* 37, 1, Article 101394 (2019), 13 pages. https://doi.org/10.1016/j.giq.2019.101394

[72] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 145–153. https://doi.org/10.1145/2872427.2883062

[73] Siobhán O'Mahony and Fabrizio Ferraro. 2007. The Emergence of Governance in an Open Source Community. *Academy of Management Journal* 50, 5 (2007), 1079–1106. https://doi.org/10.5465/amj.2007.27169153

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

[75] Elinor Ostrom. 2000. Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives* 14, 3 (September 2000), 137–158. https://doi.org/10.1257/jep.14.3.137

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

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

[78] Jessica A. Pater, Moon K. Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. ACM, New York, NY, USA, 369–374. https://doi.org/10.1145/2957276.2957297

[79] Pew Research Center. 2017, July. *Online Harassment 2017*. Report. Pew Research Center, Washington, D.C. https://www.pewinternet.org/2017/07/11/online-harassment-2017/

[80] David J Phillips. 1996. Defending the Boundaries: Identifying and Countering Threats in a Usenet Newsgroup. *The Information Society* 12, 1 (1996), 39–62. https://doi.org/10.1080/019722496129693

[81] Elizabeth Reid. 1999. Hierarchy and Power: Social Control in Cyberspace. In *Communities in Cyberspace* (1st ed.), Marc A. Smith and P. Kollock (Eds.). Routledge, New York, NY, USA, 107–134.

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

[83] Sarah Roberts. 2018. Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday* 23, 3 (2018), 9. https://doi.org/10.5210/fm.v23i3.8283

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

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

[86] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *New Media & Society* (2020), 1461444820913122. https://doi.org/10.1177/1461444820913122

[87] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. 2019. 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)*. ACM, New York, NY, USA, Article 606, 14 pages. https://doi.org/10.1145/3290605.3300836

[88] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The Social Roles of Bots: Evaluating Impact of Bots on Discussions in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 157 (Nov. 2018), 29 pages. https://doi.org/10.1145/3274426

[89] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. 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)*. ACM, New York, NY, USA, 12. https://doi.org/10.1145/3313831.3376708

[90] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443. https://doi.org/10.1177/1461444818821316

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

[92] Anna DuVal Smith. 1999. Problems of Conflict Management in Virtual Communities. In *Communities in Cyberspace* (1st ed.), Marc A Smith and P Kollock (Eds.). Routledge, New York, NY, USA, 135–166.

[93] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* 21, 9 (2019), 1910–1927. https://doi.org/10.1177/1461444819834317

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

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

[96] Allucquère Rosanne Stone. 1993. What Vampires Know: Transsubjection and Transgender in Cyberspace. Talk Given at the "In Control: Mensch-Interface-Maschine" Conference in Graz, Austria.

[97] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 1526–1543.

[98] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N. Bazarova. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 118 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359220

[99] Tiziana Terranova. 2000. Free labor: Producing culture for the digital economy. *Social text* 18, 2 (2000), 33–58.

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

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

[102] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. 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).* ACM, New York, NY, USA, 1231–1245. https://doi.org/10.1145/2998181.2998337

[103] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383. https://doi.org/10.1177/1461444818773059

[104] Kevin Wise, Brian Hamman, and Kjerstin Thorson. 2006. Moderation, Response Rate, and Message Interactivity: Features of Online Communities and Their Effects on Intent to Participate. *Journal of Computer-Mediated Communication* 12, 1 (10 2006), 24–41. https://doi.org/10.1111/j.1083-6101.2006.00313.x

[105] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, Article 160, 13 pages. https://doi.org/10.1145/3290605.3300390

[106] Scott Wright. 2006. Government-run Online Discussion Fora: Moderation, Censorship and the Shadow of Control1. *The British Journal of Politics and International Relations* 8, 4 (2006), 550–568. https://doi.org/10.1111/j.1467-856x.2006.00247.x

[107] Bingjie Yu, Katta Spiel, Joseph Seering, and Leon Watts. 2020. "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).* Association for Computing Machinery, New York, NY, USA, 253:1–9. https://doi.org/10.1145/3334480.3383009

Received January 2020; revised June 2020; accepted July 2020