Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator

SHAGUN JHAVER, Georgia Institute of Technology, USA
IRIS BIRMAN, Georgia Institute of Technology, USA
ERIC GILBERT, University of Michigan, USA
AMY BRUCKMAN, Georgia Institute of Technology, USA

What one may say on the internet is increasingly controlled by a mix of automated programs, and decisions made by paid and volunteer human moderators. On the popular social media site Reddit, moderators heavily rely on a configurable, automated program called 'Automoderator' (or 'Automod'). How do moderators use Automod? What advantages and challenges does the use of Automod present? We participated as Reddit moderators for over a year, and conducted interviews with 16 moderators to understand the use of Automod in the context of the sociotechnical system of Reddit. Our findings suggest a need for audit tools to help tune the performance of automated mechanisms, a repository for sharing tools, and improving the division of labor between human and machine decision making. We offer insights that are relevant to multiple stakeholders — creators of platforms, designers of automated regulation systems, scholars of platform governance, and content moderators.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation, automated moderation, Automod, platform governance, mixed initiative, future of work

ACM Reference Format:
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

1 INTRODUCTION

Online discussion sites provide a valuable resource for millions of users to exchange ideas and information on a variety of topics. However, the freedom these sites provide to their content creators makes them inherently difficult to govern. The utility of these sites is often undermined by the presence of various types of unwanted content such as spam, abusive, and off-topic postings. Platforms try to combat this problem by implementing processes that determine which posts to allow on the site and which to remove. We refer to the sociotechnical practices that constitute this task as "regulation mechanisms." Efficient regulation mechanisms ensure that low-quality

Authors' addresses: Shagun Jhaver, Georgia Institute of Technology, School of Interactive Computing & GVU Center, Atlanta, GA, 30308, USA, sjhaver3@gatech.edu; Iris Birman, Georgia Institute of Technology, School of Interactive Computing & GVU Center, Atlanta, GA, 30308, USA, irisb2002@gmail.com; Eric Gilbert, University of Michigan, Ann Arbor, USA, eegg@umich.edu; Amy Bruckman, Georgia Institute of Technology, School of Interactive Computing & GVU Center, Atlanta, GA, 30308, USA, asb@cc.gatech.edu.

1073-0516/2019/07-ART31 $15.00
https://doi.org/10.1145/3338243

contributions don't drown out worthy posts on the site and exhaust the limited attention of users [62]. More broadly, these mechanisms help platforms address issues like illegal content, copyright violations, terrorism and extremist content, revenge porn[1], online harassment, hate speech and disinformation [13]. It is important for platforms to have such mechanisms in place in order to protect their brand, prevent their users from attacking one another, and keep discussions productive and civil [39].

To carry out these regulation tasks, platforms rely on paid or volunteer human workers, called moderators. Although moderators on different platforms adopt differing strategies to curate their content, automated tools are increasingly playing an important role in the regulation of posts across these sites. For example, many social media websites are using machine learning tools to identify images that violate copyright law and remove them [90]. As these tools become more and more sophisticated, their role in the enactment of content regulation will likely grow over time. It is also expected that future regulatory mandates will further heighten the need for automated tools because government regulators are increasingly expecting platforms to quickly remove hate speech and other illegal content [109]. Therefore, it is critical that we understand the adoption and use of these tools in current moderation systems.

In this paper, we study the use of automated tools for moderation on Reddit, a popular discussion site [2]. Reddit adopts a "community-reliant approach" [13] to content moderation. That is, it is divided into thousands of independent communities, each having its own set of volunteer moderators, posting guidelines and regulation tools. While scholarly accounts of regulation mechanisms on discussion sites like Reddit have usually focused on how moderators collaborate with one another, create new rules, and interact with community members to enact efficient and acceptable content curation [29, 66, 75, 77], relatively little research has focused on how moderators use automated mechanisms. Although some scholars have recognized the importance of automated regulation tools [90, 109], at present, we lack a clear understanding of content regulation as a relational process that incorporates both automated tools and human labor.

In addition to filling this salient gap in research, studying automated moderation tools on Reddit is particularly important because they form a critical component of the Reddit regulation system and they perform large proportions of all regulation actions [52]. It is necessary to examine the sociotechnical practices of how human workers configure and use automated tools so that we can identify the structural challenges of current regulation systems. At a broader level, it is crucial that we understand the moderation apparatus that social media companies have built over the past decade if we hope to provide practical solutions to the hard questions being asked right now about how regulation systems can distinguish freedom of expression from online harassment [53], or how they can make decisions on content that may be too graphic but is newsworthy, educational, or historically relevant [40].

Unfortunately, given the black-boxed nature of many social media platforms, the process of content regulation remains opaque on many sites. That is, it is hard for outsiders to infer behind-the-scenes operations that guide it. Reddit, however, provides an excellent opportunity to study the sociotechnical details of how automation affects regulation processes because its moderators are *not* paid employees bound by legal obligations to conceal the details of their work but are volunteer users. Taking advantage of this opportunity, we conducted interviews with sixteen Reddit moderators, and analyzed the ways in which the use of automated tools reshapes how Reddit moderators conduct their work. In doing so, we offer insights into the tradeoffs that arise because

[1]Revenge porn or "nonconsensual pornography" involves "the distribution of sexually graphic images of individuals without their consent" [23]. This includes images originally obtained *with* consent, usually by a previously intimate partner [22], as well as images obtained *without* consent, e.g., through hidden recordings [23].

of the use of automated tools, the tensions of redistributing work between automated tools and human workers, and the ways in which platforms can better prepare themselves to adopt these tools.

We explore the following research questions in this paper:

(1) How are automated tools used to help enact content regulation on Reddit?
(2) How does the use of automated tools affect the sociotechnical process of regulating Reddit content?
(3) What are the benefits and challenges of using automated tools for content regulation on Reddit?

While preparing for this study, we found that Reddit has an open-access API (Application Programming Interface) that allows bot developers to build, deploy and test automated regulation solutions at the community level. Access to this API has encouraged the creation and implementation of a variety of creative automated solutions that address the unique regulation requirements of different Reddit communities. We focused on one of the most popular automated tools, called Automoderator (or Automod), that is now offered to all Reddit moderators. Automod allows moderators to configure syntactic rules in YAML format (Figure 2) so that these rules make moderation decisions based on the configured criteria. We show that Automod not only reduces the time-consuming work and emotional labor required of human moderators by removing large volumes of inappropriate content, it also serves an educational role for end-users by providing explanations for content removals.

Despite the many benefits of using Automod, its use also presents certain challenges. Prior CSCW research has established that a fundamental social-technical gap exists between how individuals manage information in everyday social situations versus how this is done explicitly through the use of technology. Often, technical systems fail to provide the flexibility or ambiguity that is inherent in normal social conditions [1]. In line with this, our findings reveal the deficiencies of Automod in making decisions that require it to be attuned to the sensitivities in cultural context or to the differences in linguistic cues.

Building on our case study of Reddit Automod, we provide insights into the challenges that community managers can expect to face as they adopt novel automated solutions to help regulate the postings of their users. For example, they may have a reduced level of control over how the regulation system works — as moderators reduce the number of posts that they manually review and delegate to automated tools, these tools may make mistakes that could have been avoided owing to their limitations of evaluating the contextual details. Moreover, moderators may not be able to understand the reasons behind some actions taken by automated tools. Another possible challenge is that moderators may have to make decisions about the levels of transparency they show in the operation of automated tools — if they are too transparent about how these tools are configured, these tools may be exploited by bad actors.

In addition to these challenges, we also highlight how the use of automated tools may affect how moderators design community guidelines. We found that Reddit moderators sometimes create posting guidelines that play to the strengths of Automod so as to make the work of moderation easier. For example, guidelines like "describe the image you're posting" provide additional material for Automod to catch. However, complying with such guidelines may increase the amount of work that end-users have to perform. Therefore, using automated tools may affect not only the moderators but also the other stakeholders in content regulation systems.

There is a growing enthusiasm among many companies hosting user contributions to use machine learning and deep-learning-based tools to implement content regulation and relieve human moderators [13, 71, 83]. While the accuracy of such tools has risen for many kinds of

moderation tasks, the tools often can't explain their decisions, which makes mixed-initiative human-machine solutions challenging to design. Human-understandable ML is an active area of research [64]. Yet, as we will see, Automod does not rely on machine learning techniques but it rather uses simple rules and regular-expression matching which can be understood by technically savvy human moderators. We found that moderators self-assess their skills at configuring Automod, practice care when editing Automod rules, and coordinate with other moderators through external channels like Slack to resolve disagreements. Thus, Reddit moderation is an intriguing example of human-machine partnership on a complex task that requires both rote work and nuanced judgment.

We organize the remainder of this article as follows: we start by providing a brief overview of the sociotechnical system of Reddit moderation. Next, we examine prior research on content regulation and automated regulation mechanisms. Following this, we present our methods of data collection and analysis, and describe our participant sample. Next, we discuss our findings on the development and use of automated moderation tools, focusing on Reddit Automod as a case study. We then discuss the limitations of currently used automated techniques and the challenges they pose for Reddit moderators, emphasizing insights that may be useful for other platforms as they adopt automated regulation tools. We conclude with highlighting the takeaways of this research for different stakeholders.

2 STUDY CONTEXT: REDDIT MODERATION

Reddit is composed of thousands of small and large communities called subreddits where users can post submissions or comment on others' submissions. Activity on Reddit is guided by a user agreement[2] and content policy[3] similar to the terms and conditions of many websites and a set of established rules defined by Reddit guidelines called Reddiquette[4] [32]. Each subreddit also has its own set of rules that exist alongside site-wide policy and lay out what content is acceptable and what is not acceptable on that subreddit. These rules are typically found in sidebars of the subreddit [74]. They have higher salience than Reddiquette for most users [62]. Many subreddits have a separate set of rules for submissions and comments. Content removals occur frequently on Reddit. Analyzing Reddit submissions for a separate project [52] using data collected from the pushshift.io service[5], we found that 21.77% of all submissions posted on Reddit between March - October 2018 (79.92 millions submissions posted; 17.40 million submissions removed) were removed.

Reddit moderators are volunteer Reddit users who take on the responsibility of maintaining their communities by participating in a variety of tasks. These tasks include (1) coordinating with one another to determine policies and policy changes that guide moderation decisions, (2) checking submissions, threads and content flagged by users for rule violations, (3) replying to user inquiries and complaints [6], (4) recruiting new moderators, (5) inviting high-profile individuals to conduct AMA (Ask Me Anything) sessions [80], (6) creating bots [70] or editing Automod rules (described below in this section) to help automate moderation tasks, and (7) improving the design of the subreddit using CSS tools. Moderators usually prefer to focus primarily on a few of these task categories depending on their interests, skills, prior moderation experience, level of access[7],

[2] https://www.reddit.com/help/useragreement/
[3] https://www.reddit.com/help/contentpolicy/
[4] https://www.reddit.com/wiki/reddiquette
[5] https://files.pushshift.io/reddit/submissions/
[6] Users can reach out to moderators using ModMail, a personal messaging tool that forwards inquiries to all moderators of a subreddit.
[7] Moderators can be given different levels of access on a subreddit depending on their roles. Different binary flags can be set to provide different permissions. For example, 'access' flag allows moderators to manage the lists of approved submitters and banned users, and 'posts' flag allows moderators to approve or remove content. Only moderators with 'full permissions' can change the permission levels for other moderators.

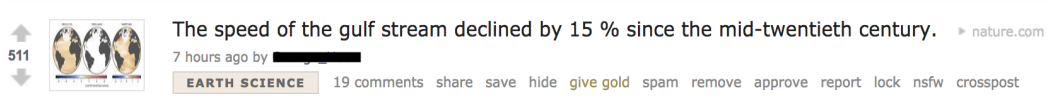

Fig. 1. Moderators' view of a submission post. This interface differs from a regular user's interface because of the presence of additional links for 'spam', 'remove', 'approve', 'lock' and 'nsfw'. Moderators can use these links to take corresponding actions on the post. For example, clicking 'lock' prevents the submission post from receiving any new comments.

influence among the moderators and the requirements of the subreddit. For the purpose of this paper, we focus on moderation activities that are related to removing content using automated mechanisms.

One automated solution to content regulation is "Automoderator" (or "Automod"), which first became popular as a third-party bot but was later incorporated into Reddit and offered to all the moderators. Many moderators use Automod to help improve their work efficiency. This solution uses a filtering approach where moderators codify phrases that usually occur in undesirable posts as regular expressions[8] (also called 'regex') into a wiki which they regularly update. Automod then scans each posted material for the presence of the coded phrases and removes the material if it contains any such phrase. In this paper, we investigate the use of Automod tool as a component of the sociotechnical system of content regulation. We also briefly discuss how other bots are used to improve the efficiency of moderation tasks. Our findings highlight how the currently employed human-machine collaborations help manage content quality on Reddit.

Reddit provides its moderators with alternate interfaces that contain tools that are not accessible to regular users of the subreddit (see Figure 1). These tools can be used to perform a variety of moderation actions. For example, for each submission, moderators can remove the submission thread, lock the thread from receiving any new comments, or remove any comment on the submission thread. Each subreddit also has its own private moderation log, a tool that allows moderators to view which posts and comments have been removed, at what time, and by which moderator. Although Reddit provides these moderation tools to all moderators by default, many moderators find these tools inefficient and cumbersome to use. This has motivated the creation of third-party front-end tools that help moderators regulate their communities more efficiently. We discuss how such tools and back-end bots are built and shared between different communities in our findings (Section 5.1).

3 RELATED WORK

We now situate our research amid prior scholarship on content regulation and automated regulation mechanisms. We also discuss how we contribute to a growing body of literature that focuses on Reddit moderation.

3.1 Content Regulation

Internet platforms that are supported by advertising as their business model have an economic incentive to maximize site activity to increase ad revenue. They have an interest in adopting a rhetoric of free speech advocacy and allowing users to post as much content as possible [109]. At the same time, they have competing incentives to remove material that is likely to make their users uncomfortable [93], such as obscene or violent material, because keeping such content risks

[8]Regular expressions are special text strings for describing specific search patterns. They are used to search a volume of text for a group of words that match the given patterns [107].

having users leave the platforms [63]. Consequently, platforms have to address the challenges of exactly when to intervene and in what ways and negotiate competing imperatives to keep as much content as possible online while removing material that could alienate users and advertisers [109]. As platforms grow, they face the challenge of scale in their role as content curators [41, 71]. For example, many large Reddit communities have millions of subscribers, and their regulation systems have to process thousands of new posts every day.

To address the challenges of processing high volumes of content efficiently, social media companies have created intricate and complex systems for regulating content at scale [44, 106]. Such systems usually consist of a small group of full-time employees at the top who set the rules and oversee their enforcement, adjudicate hard cases, and influence the philosophical approach that the platforms take to govern themselves [38]. Platforms also employ a larger group of freelancers who work on contracts with them and guard against infractions on the front line [20, 91]. Finally, platforms rely on regular users to flag content that offends readers or violates the community rules [26]. Many platforms also depend on users to evaluate the worth of each comment. This is done by aggregating how users rate that comment. This process is often referred to as distributed content moderation, and its benefits and limitations have been studied through prior research on moderation of the Slashdot website [65, 66]. This research shows that distributed moderation can enable civil participation on online forums [67].

While the use of commercial content moderators is popular in the industry, not all platforms use paid staff to moderate their content. Many online platforms (e.g., Reddit, Wikipedia, Facebook Groups) largely rely on volunteer moderators who are given limited administrative power to remove unacceptable content and ban problematic users [75, 77]. These moderators are typically selected from among the users who are most actively involved in the community and are invested in its success [75]. Thus, they are well-suited to understand the local social norms and mores of the community [29]. They create and enforce local rules that establish the grounds for acceptable content. Although these volunteer moderators are not employees of the platforms, as communities get larger, moderators are often viewed by the users as representatives of the platforms. Therefore, moderators constantly negotiate their positions with communities as well as with platforms [75].

Content regulation is not only a difficult but also an important task. It helps internet platforms present their best face to new users, advertisers, investors and the public at large [39]. Having taken on a curatorial role, these platforms "serve as setters of norms, interpreters of laws, arbiters of taste, adjudicators of disputes, and enforcers of whatever rules they choose to establish" [39]. Because the moderators[9] who help achieve content curation are responsible for adjudicating what millions of users see and, just as importantly, what they don't see, they play an important role in modern free speech and democratic culture. Indeed, the impacts of content regulation transcend online experiences and the question of freedom of speech [53, 109]. Prior research has suggested that failures in content regulation may cause irreversible professional damage [55, 109] or they may result in disabled or elderly users losing the ability to communicate with their family networks [109]. Therefore, it is important to study the processes that allow moderators to make complex contextual choices and enact local social norms.

Some prior research has explored the work of moderators [5, 29, 31, 69, 97, 111]. For example, Epstein and Leshed studied how moderators facilitate public deliberations on RegulationRoom, an online platform for policy discussions, and found a strong need to automate more of the moderators' work [31]. Diakopoulos and Naaman studied regulation of online news comments on the website

[9]Note that although regular users also contribute to content moderation by using flagging and reporting mechanisms, for the rest of this paper, we use the term 'moderators' to refer to volunteer moderators with administrative power to manage content and configure the communities they moderate.

SacBee.com [29]. They found that although the moderators were concerned about handling a large volume of comments on the site, they were reluctant to outsource the control of content moderation. Moderators felt that outsiders might not have a locally meaningful understanding of the issues to be able to make appropriate moderation decisions in hard cases [29]. We explore how a similar inclination to maintain control affects the acceptance and use of Automod among Reddit moderators. Most recently, Seering et al. studied volunteer moderators on Twitch, Reddit and Facebook, and presented a systematic description of the full process of governance across these multiple platforms [97]. We add to this rich body of research by focusing on moderators of a single platform, Reddit, and their activities in a specific context—the interplay with automated tools, aiming to glean transferable insights for moderation on all platforms.

A complementary line of research has focused on understanding the "emotional labor" [48] of moderation work [30, 60, 61, 75, 77, 90, 91]. Sarah T. Roberts studied commercial content moderation [90–92] and found that many social media companies use the services of workers who are dispersed globally. She pointed out that the routine, factory-like nature of the work of content moderation leads to burnout among many workers. Roberts also noted that the constant viewing of troubling content takes an emotional toll on the workers, and they resist discussing their work with friends and family to avoid burdening them [90]. In a similar vein, Kerr and Kelleher pointed out that such workers have to perform the emotional labor of enacting an "apolitical, culturally nuanced subjectivity" in their online work that may not align with their offline identity [60]. Several recent media articles have also highlighted the emotional labor of moderation work [11, 12, 34, 89, 104]. We add to this literature by showing how automated content regulation on Reddit helps reduce the workload of moderators and allows them to avoid viewing large volumes of offensive content. We also discuss the trade offs of this reduction in emotional labor via the use of automated tools, e.g., we discuss the problem of mistaken automated removal of posts that may contain offensive language but are appropriate for the community in the context of their discussion.

A few researchers have focused on understanding content moderation on Reddit [17, 18, 32, 75, 77]. Fiesler et al. conducted a mixed-methods study of 100,000 subreddits and contributed a comprehensive description of the type of rules that are enacted across Reddit [32]. Matias showed how Reddit moderators respond to users' complaints of censorship [75]. Kiene et al. studied the role of moderators in welcoming newcomers to rapidly growing subreddits [61]. McGillicuddy et al. studied the political and moral struggles that Reddit moderators face in their work [77]. They emphasized the moderators' awareness of their power over the community. We contribute to this growing body of research on Reddit moderation by highlighting the role of technology in assisting the work of content moderators. We distinguish our work from prior research on social dynamics of Reddit moderators (e.g., [73, 75, 77]) by bringing to scrutiny the moderator interactions that are related to the use of Automod. We also build upon our findings to discuss the design implications for building mixed-initiative content regulation systems.

3.2 Automating Content Regulation

Regulating content usually involves making trade-offs between the goals of achieving high efficiency and sustaining low costs [44]. It is possible that regulation systems can considerably prevent spam, harassment and other forms of abuse on a large community if enough number of expert human moderators are available to carefully review each post. But that may drive up the costs of moderation to unacceptable levels. One potential solution to minimize such costs is to automate content regulation. Recognizing this, some scholars have begun developing automated (often, machine learning based) solutions to automate certain aspects of content regulation [28, 42, 45, 68, 94, 108]. For example, researchers have proposed computational approaches to identify hate speech [18], pornography [98] and pro-eating disorder content [15]. Wulczyn et al. created a machine learning

classifier trained on human-annotated data to identify personal attacks in online discussions on Wikipedia [112]. Park et al. designed a visual analytic tool, CommentIQ, that can help moderators select high-quality comments on online news sites at scale [85].

Alongside the growing scholarly interest in automated moderation, many platforms are also increasingly deploying tools that automate the process of identifying problematic material and taking appropriate moderation decisions [71]. For example, The Washington Post uses ModBot, a software application that employs NLP and machine learning techniques, to automatically moderate user-comments on news articles [58]. Although not machine-learning based, Reddit Automoderator is an excellent example of automated systems that are widely embraced by online communities [78, 89]. There is optimism in the industry that AI and machine learning systems can eventually replace the thousands of human workers who are currently involved in making moderation decisions, either voluntarily or as paid workers [71]. Such systems also have the potential to execute moderation in a much faster way than human moderators.

Despite the enthusiasm for automated moderation systems among social media platforms as well as researchers, such systems face many challenges. Critics have argued that the currently available AI technologies are not good at understanding the context of a given post, user or community [71]. Therefore, they may end up resulting in many false positives[10], that is, posts that are not problematic to the community get removed. Worse still, Blackwell et al. found that using automated approaches to identify abusive language can result in situations where the concerns of only the dominant groups are emphasized and existing structural inequalities are exacerbated [8]. These systems are also vulnerable to the same challenges that human moderators face – many moderation decisions are complex, subtle and contested, and different users may feel differently about the appropriateness of the same content [39, 53].

Moderators of many subreddits configure and manage the Automod tool (see Section 2) as part of content regulation [81]. While this solution is not machine learning (ML)-based, it still automates moderation for a large number of posts and comments. In this paper, we investigate how moderators use this tool and the benefits and limitations of using this tool. We study content regulation on Reddit as a sociotechnical system [82]. This allows us to attend to the complexities generated by the coupling between technical and social components of this system [84]. Through illuminating the impact of automation on moderators and considering how these workers will need to develop new skills in order to use more advanced moderation tools, our research adds to a broader literature [49, 54, 56] on the future of work at the human-technology frontier [33].

Our work is related to Geiger and Ribes' research on the use of software tools to enforce policies and standards on Wikipedia [36]. They showed that bots on Wikipedia automatically revert edits to its pages based on "criteria such as obscenity, patent nonsense, mass removal of content, and various metrics regarding the user who made the edit." Geiger and Ribes also found that assisted editing programs show new edits to Wikipedia editors in queues and allow them to quickly perform actions such as reverting edits or leaving a warning for the offending editor [36]. Our work adds to this literature by describing the coordinated actions of moderators and automated tools in the context of Reddit. Although there are substantial differences in design mechanisms, outputs of content regulation, and functioning of software tools between Wikipedia and Reddit, Geiger and Ribes' work [36] provides an excellent point of comparison to our research on how human and non-human work can be combined to facilitate content regulation.

[10]We call a post a true positive if the regulation system removes that post and the moderators consider that post removal appropriate. Although it might seem counter-intuitive to use the term 'true positive' to denote a correctly *removed* post, it highlights the focus of Reddit regulation system on removal of inappropriate content. Besides, this is in line with our participants' use of the term 'true positive' to refer to correct removals.

Table 1. Activity level, date of creation and number of moderators for sampled subreddits

Subreddit	Total comments	Creation date	# Moderators
oddlysatisfying	180,978	May 15, 2013	22
politics	14,391,594	Aug 06, 2007	37
explainlikeimfive	964,821	Jul 28, 2011	38
space	795,186	Jan 26, 2008	23
photoshopbattles	300,369	Jan 19, 2012	23

4 METHODS

This study was approved by the Georgia Institute of Technology IRB. The study included 16 in-depth, semi-structured interviews with Reddit moderators. Next, we provide the details of our study design.

4.1 Selection of Subreddits

Prior research suggests that as communities grow, human-only moderation becomes increasingly unfeasible, and the dependency on automated tools increases [7, 38, 90]. The first and fourth authors' long-term experiences as moderators on many large and small subreddits also indicate that moderation becomes more demanding, complicated and involved as subreddits grow large. Therefore, we decided to focus on large subreddits, aiming to unpack how automated mechanisms help regulate large, active, long-running subreddits. To that end, we used a purposive sampling approach to select participants for our study [79]. The power of this sampling lies in selecting information-rich cases whose study illuminates the questions of interest [86]. We used this approach to recruit participants who moderate large, high-traffic subreddits.

We began with a list of the 100 largest subreddits, as determined by their subscriber count, available on the RedditMetrics website [11]. We sampled subreddits from this list that are at least five years old, show high levels of activity (at least 100,000 comments posted in the period June 1 to Dec 31, 2017[12]) and reflect a diverse range of topics and moderation rules. Our sampled subreddits include r/photoshopbattles, r/space, r/explainlikeimfive, r/oddlysatisfying and r/politics (Table 1). We received permission from our participants to disclose the names of these subreddits. We chose not to anonymyze the names of these subreddits because knowing the identity of these subreddits is important to ground our research and help readers contextualize the nuances of our findings.

We observed about a hundred posts on each of these five subreddits and found that these communities reflect a wide range in the themes of conversations. These subreddits show some overlap in their submission guidelines (e.g., all five subreddits ask users to "be civil" or "be nice"). However, most guidelines on these subreddits directly reflect the focus and norms of the communities and are therefore quite unique to them. For example, r/photoshopbattles has a list of seven rules about the types of images that are allowed to be submitted, reflecting the focus of the subreddit on image submissions. To take another example, a majority of rules on r/explainlikeimfive focuses on the types of questions and explanations that are allowed to be posted on the subreddit. These subreddits also differ in how sensitive and emotionally charged the discussions are. Therefore, selecting these subreddits allowed us to gain insights into a diverse range of moderation practices that are related to the use of automated mechanisms.

[11] http://redditmetrics.com/top

[12] We retrieved this data by running SQL-like database queries on the public archives of Reddit dataset hosted on the Google BigQuery platform [6].

4.2 Interviews

We interviewed three moderators from each of the five subreddits selected in our sample so that we could triangulate themes from multiple perspectives on the work of moderation for each subreddit and attain a deeper understanding. In addition, we interviewed Chad Birch, a Reddit moderator (past moderator of r/games; currently moderates r/automoderator) who created Automoderator. We invited moderators to participate in semi-structured interviews with us by contacting them through Reddit mail. We also met many Reddit moderators at CivilServant Community Research Summit, a gathering of moderators and researchers held at the MIT Media Lab in Boston in January 2018, and recruited some moderators present who moderated any of the five subreddits in our sample. Participation was voluntary and we did not offer any incentives.

In our interviews, to understand the role of Automod in context, we first asked participants more general questions, such as how they became moderators and what typical tasks they engage in while moderating. Next, we asked them questions about how they use automated moderation mechanisms on their subreddit. We inquired about the type of conversations they try to foster in their communities and how automated mechanisms help them attain their goals. We also discussed with our interviewees how they coordinate with other moderators to configure automated regulation tools and how the subreddit policies affect their use of these tools. Finally, we asked participants about the limitations of automated tools and the challenges they face in using them.

Each interview session lasted between 30 and 90 minutes, and was conducted over the phone, on Skype, or through chat. We contacted some participants for further clarification of their responses during the analysis stage. Although we attempted to get additional data such as Automod configuration rules and moderation log (described in Section 2), our participants were hesitant in providing us these data because they contain sensitive information. Even so, some moderators calculated and sent us the proportions of moderation work done by Automod in their subreddit (Table 3). Importantly, in their interviews with us, all our participants were forthright and provided us many specific examples of how they use Automod and how they enforce different subreddit rules. Therefore, we have primarily relied on our interview data to present our findings.

The first and fourth authors of this study collectively spent over 100 hours moderating multiple large (e.g., r/science) and small (e.g., r/youtubers) subreddits over the last year to understand the dynamics of Reddit moderation systematically. We supplemented our interview data with participant-observation field notes [105] taken while moderating these subreddits.

4.3 Participants

Sixteen moderators participated in our study. All of our participants were male[13]. Fifteen participants reported being in their 20s or 30s, and one participant chose not to share his age. Although a majority of our participants are from the US, we also interviewed moderators from UK, Ireland, Canada, Australia, and India. Table 2 provides some demographic and moderation experience related information about our participants. We use light disguise [9] in describing our participants in this table and in our findings. Therefore, although we have omitted sensitive details to protect the identity of our participants, some active members of their communities may be able to guess who is being discussed. We also note that we have not anonymized Chad Birch, the creator of Automod, after asking for his permission and to provide him credit for his work [10].

We note that our approach to subreddit selection introduced some limitations in our study. Specifically, our results are based only on participants who moderate five large subreddits (in addition to the creator of Automod). Although our participants moderate a variety of large as well as small subreddits, our interviews mostly focused on moderation of large subreddits. Therefore,

[13]We discuss this limitation in Section 6.4.

Table 2. Study Participants. This table provides the following information about our participants: the subreddit in our sample that they moderate, their age, total number of subreddits that they currently moderate, and their tenure as a moderator on the corresponding subreddit. If participants did not want to share certain information or the data was unavailable, we have noted it with "NA".

Subreddit	Participant	Age	# of subs moderated	Tenure
r/photoshopbattles	PB_1	33	91	2 years
r/photoshopbattles	PB_2	Late 20's	5	4 years
r/photoshopbattles	PB_3	NA	7	5 years
r/space	$Space_1$	25	2	1 year
r/space	$Space_2$	20-25	NA	NA
r/space	$Space_3$	33	11	1 year
r/oddlysatisfying	OS_1	32	28	4 years
r/oddlysatisfying	OS_2	21	12	10 months
r/oddlysatisfying	OS_3	26	8	3 years
r/explainlikeimfive	$ELIF_1$	30	8	5 years
r/explainlikeimfive	$ELIF_2$	27	8	4 years
r/explainlikeimfive	$ELIF_3$	28	3	1 year
r/politics	Pol_1	32	8	3 years
r/politics	Pol_2	25	12	1 year
r/politics	Pol_3	25-29	5	1 year
r/Automoderator	Chad	34	8	6 years

our findings should be interpreted as representative of moderation on large communities only. Even though we focused on only five Reddit communities, conducting independent interviews with three moderators from each community allowed us to check our participants' interpretations of events and processes against alternative explanations. It also helped us discover the properties and dimensional ranges of relevant concepts in our analysis [102]. We also stress that our participants were quite diverse in terms of their backgrounds, including their tenure as a moderator and the number and types of communities they moderate.

4.4 Analysis

We fully transcribed the data from the interviews and read it multiple times. Next, we applied interpretive qualitative analysis to all interview transcripts and field notes [79]. This process entailed a rigorous categorization of data as we identified relevant patterns and grouped them into appropriate themes. Our analysis began with "open coding" [19], in which we manually assigned short phrases as codes to our data. This first round of coding was done on a line-by-line basis so that codes stayed close to data. We gathered 481 first-level codes at this stage. Examples of first-level codes include "40-50% of moderation action taken by Automod" and "Automod rule resulting in many mistaken removals."

Next, we conducted multiple subsequent rounds of coding and memo-writing. We engaged in the continual comparison of codes and their associated data with one another. All the authors discussed the codes and emerging concepts throughout the analysis. After the first round of coding that closely followed the text, our next round of coding was more high level and resulted in codes such as "Refining Automod rules over time" and "Finding Automod to have a steep learning curve."

In subsequent rounds of coding, we combined and distilled our codes into seven key themes. These themes included "Use of automated moderation tools other than Automod" (discussed in Section 5.1), "Automod creation and incorporation into Reddit" (Section 5.2.1), "Utility of Automod" (Section 5.2.2), "Use of Automod to enforce community guidelines" (Section 5.3), "Social dynamics around the use of Automod" (Section 5.4.1), "Configuring Automod rules" (Section 5.4.2) and

"Challenges of using Automod" (Section 5.5). In addition to the ones reported in this paper, themes such as "Becoming/continuing to be a moderator" and "Recruiting new moderators" emerged but were excluded in further analysis. Next, we present our findings.

5 FINDINGS

We now present our findings from our interviews with Reddit moderators. We begin with a discussion of how Reddit moderators build, use and share a variety of automated tools. Following this, we focus on the creation and use of Automod. First, we showcase the creation and incorporation of Automod into Reddit, and highlight the utility of Automod for moderation teams. Next, we outline the use of Automod for enforcing different types of community guidelines. We then present our findings on the mechanics of Automod configuration. Finally, we discuss how Automod creates new tasks for Reddit moderators.

5.1 Reddit Moderation Tools

Reddit moderators use a variety of moderation bots to automate removal of undesirable content. Our participants told us that Reddit has an open and easy-to-use API that promotes the development of such bots. One popular bot checks for whether a submitter has 'flaired'[14] a post. Another popular bot called 'Botbust' identifies and bans Reddit bots that post spam or offensive content or comments that provide no value to the community [14]. Yet another bot helps moderators use their phones for moderating tasks by looking out for specially formatted comment tags left by them.

> *"What they'll do for example is, they'll have a bot that looks for a moderator leaving a comment like, '!R1.' If it sees a comment like that made by a moderator, that means, remove this post for violating rule one... It lets the moderators do this big process of removing the post, leaving an explanation comment, everything like that that would be very tedious to do on mobile manually, just automatically by leaving a really short comment." - Chad*

We found that some moderators prefer to design and implement moderation bots from scratch. Five participants told us that certain moderators on their subreddits create bots themselves so that they can attend to the specific needs of the community. Pol_1 pointed out that many subreddits even recruit users who are adept at writing Reddit bots as moderators so that those users can help automate various content regulation tasks. In a similar vein, PB_2 noted:

> *"We have multiple bots that we have made ourselves to automate out tasks such as catch[ing] plagiarism and detect[ing] reposts."*

While some moderators build bots from scratch, others frequently use tools made by other subreddits. Reddit moderators often moderate multiple subreddits (see Table 2) and develop relationships with moderators of many communities. Our participants told us that they use their connections with moderators from other subreddits to borrow tools and bots that improve content regulation for their own communities. For example, PB_2 said that r/photoshopbattles only allows submission of images with reasonable quality but the community did not have any tools to automatically check image attributes like image resolution. In order to automate checking image quality, the moderators of r/photoshopbattles borrowed a bot from another subreddit and deployed it on their subreddit. Similarly, OS_3 said:

[14]Flairs are usually used to arrange submissions into different categories. For example, r/science is a popular subreddit where users post links to peer-reviewed research. Submitters are required to flair their posts by indicating the subject of the research, e.g., Chemistry, Astronomy, Paleontology, etc. Flairs allow readers to quickly filter for posts that they are interested in.

> *"We are talking about using one of /r/technology bots that will help reduce the amount of spam that we get. This was someone's pet project so it has been tailored to specifically what we want."*

This sharing of moderation tools indicates that social interactions between moderators of different communities play a role in helping moderators discover and implement automated mechanisms for improving content regulation. Indeed, lack of such informal interactions may lead to duplication of efforts in creating moderation bots. Some participants told us that even though bots with similar functionality may have been developed by moderators of other subreddits, a lack of central repository of such bots forces them to develop their own tools from scratch.

> *"There is no binder with 'Oh, you want to do this? Here's the code for that!' So, you will often get duplicate efforts." – Pol_2*

A complementary set of regulation tools includes those that don't remove the content themselves but help moderators enact regulation tasks more efficiently. For example, a majority of our participants use Reddit toolbox [15], a browser add-on that provides many useful features such as allowing moderators to remove a comment and all its replies with a single click, and tagging pre-written reasons for removal when they remove a comment or post. Similar to many other Reddit moderation bots, Reddit toolbox is also a product of voluntary work of users dedicated to maintaining Reddit communities.

In summary, moderators on each subreddit use a wide variety of automated tools, largely developed by volunteer Redditors, to enact content regulation. Given that different communities have different moderation needs and require specific solutions, it is valuable for Reddit to have a third-party API that developers can use to build customized tools. Next, we turn to our findings on the introduction and use of Automoderator (or Automod). We dedicate the rest of our findings to discussing this tool because it is the most widely used automated regulation tool on Reddit, and it immensely affects Reddit moderation.

5.2 Introduction of Automod on Reddit

All subreddits in our sample allow every post to pass through by default and only remove a post later if a moderator wishes to reject it. Although moderators have the capability to configure settings so that only posts that get approved by the moderators appear on the site, this arrangement is unsustainable on communities that experience heavy traffic.

> *"I don't know of any big subreddits that do it because it would become untenable to physically moderate every single post before it's seen." – Pol_2*

But allowing all posts to appear by default creates a potential for situations where undesirable content is not removed. Many subreddits have a limited number of human moderators who only moderate posts at certain times of the day. Before tools like Automod were available, subreddits often had posts that were offensive or that violated the community rules, but they remained on the site for hours despite many user complaints until a human moderator accessed Reddit and noticed them. This lack of moderation often agitated the regulars of subreddits. Thus, there was a need to create an automated tool that would remove at least the most egregious postings without human intervention.

5.2.1 Automod Creation and Incorporation into Reddit. The original version of Automod was voluntarily created by Chad Birch using Reddit API [16] in January 2012. Chad was inspired to create this tool when he was a moderator on the *r/gaming* subreddit. He noticed that many of the tasks

[15] https://www.reddit.com/r/toolbox/
[16] https://www.reddit.com/dev/api/

```
---
  #Remove comments for users with comment karma lower than 1
  type: comment
  author:
    flair_text (regex): "^$"
    comment_karma: "< 1"
    is_submitter: false
  action: spam
  action_reason: Low Karma

---
```

Fig. 2. An Automod configuration snippet written in YAML format. This rule labels all comments posted by users having less than 1 comment karma score as spam and removes those comments, assigning 'Low Karma' as the removal reason. Exceptions include cases where the comment author is flaired (user flairs are digital markers usually assigned by moderators to trusted users) or where the user comments in a thread submitted by herself.

he did as a moderator were mechanical, e.g., checking the domain name of submitted posts to see whether they belonged to any of the common suspicious sites, checking whether the post submitter was a known bad actor, and looking out for some keywords that indicated that the post should be removed. He felt that such tasks were amenable to be performed automatically. Following this, he built the original Automod as a bot that could be set up with conditional checks, apply these defined checks to all newly posted content and perform the configured actions such as post removal and user ban if the checks were met [27]. These checks could be defined using regular expressions, which allowed for defining patterns in addition to specific words. For example, one subreddit configured Automod using the following regular expression to catch and remove many homophobic slurs:

(ph|f)agg?s?([e0aio]ts?|oted|otry)

This single expression catches many slur words such as 'phagot,' 'faggotry,' 'phaggoted,' etc.

These check conditions could be combined in any manner and could also be inverted so that any content not satisfying the condition could be approved or removed. Using these basic building blocks, Automod could be used to develop a variety of capabilities (e.g., see Figure 2) such as banning posts from suspicious domains, auto-approving submissions from users whose account age and karma points [17] are higher than some threshold values, and removing user-reported comments containing certain phrases.

Chad found that a significant amount of moderation work could be handled using these basic configurations. He observed that implementing this tool in the r/gaming subreddit drastically reduced the amount of human moderation needed to regulate that subreddit. Seeing this, he offered the use of Automod to various other subreddits [27]. Thereafter, Automod quickly became popular on many communities.

Eventually, Automod was officially adopted by Reddit in March 2015 and offered to all the subreddits. Currently, each subreddit has its own wiki page for configuring Automod rules [3]. This page is accessible only to the moderators of that subreddit. Moderators can define the rules for their subreddit on this page in YAML format [4], and these rules go into effect immediately after they are configured. Figure 2 shows an example of Automod rule that instantly removes comments

[17] Reddit karma are digital reward points that users gain by posting popular content.

Table 3. Automod removal rate - % of removed comments (and submissions) that were removed by Automod over a month. These values were reported by our participants.

Subreddit	For comments	For submissions
r/oddlysatisfying	29.28%	4.95%
r/photoshopbattles	81%	66%
r/politics	79.86%	33.66%
r/explainlikeimfive	57.89%	72.78%

posted by users with low karma. Reddit's official adoption of Automod further contributed to its popularity.

> *"When it [Automod] became an option on the site for the subreddit settings, it was infinitely easier to set up, and it became in-house, so it was a little bit more reliable." – Pol_2*

> *"I think, putting it in control of people directly made a big difference... they feel a lot better being able to know, 'Okay, I have this configuration. If it starts doing things that I don't want it to, I can just wipe this page out and it'll stop.' " – Chad*

Similarly, many other participants told us that they value the level of control that Automod provides them. Even when Automod mistakenly removes posts, they can usually recognize the syntactic setting that caused that removal, and change that setting to avoid similar mistakes in the future. In contrast, more advanced automated systems that use machine learning or neural nets may not be able to provide such specific causal understanding of their decisions to the moderators.

The popularity of Automod and its eventual official adoption by Reddit highlights the value of moderation tools developed by volunteers users. Next, we discuss how the introduction of Automod helped improve the efficiency of content regulation.

5.2.2 Utility of Automod. Automod can be configured to handle removal of undesirable submissions and comments separately. We discuss the automated removal of submissions and comments together in this paper. This is done because our analysis suggests that they are both configured using similar regex patterns and they present similar benefits and challenges.

After Automod was introduced, moderators were able to enforce their subreddit rules more effectively and efficiently. For example, many moderators configured rules that auto-removed posts which received more than a certain threshold of user complaints. This allowed the moderators to regulate their subreddit even when human moderators were not proactively monitoring new posts and comments throughout the day.

> *"Reddit moves so quickly that once a post is a day old, it is just irrelevant to even moderate it at that point. It was really nice to have that automated in some ways... [*Automod*] allowed certain things like strict title formatting and stuff to become possible because before, a subreddit could never really have very strict title requirements or anything just because you would need a moderator to manually enforce those title requirements and it would never happen." – Chad*

All of our participants told us that they consider Automod an indispensable tool in their work, especially for moderating large subreddits. Table 3 shows the proportion of all removed submissions and comments that were removed by Automod on different subreddits over the period of a month, as reported by our participants. Although we couldn't obtain this breakdown between submissions and comments for the r/space subreddit, $Space_3$ told us that Automod does 40-50% of all moderation actions on the r/space subreddit. $ELIF_1$ pointed out:

> *"Extensive* Automod *rules is the only reason it's possible to moderate ELIF with the number of people [human moderators] we have been able to get to help us."*

Similarly, Pol_2 said:

> *"It's so powerful! I mean, most subreddits have thousands of lines of this code that takes a lot of the menial work out of it ... If it was to go away at some point, subreddits would become horribly moderated, and basic things would just grind to a halt."*

Some participants noted that since many inappropriate postings on Reddit are made by new users who may not be aware of the rules and norms of the community, a key advantage of Automod is that it can be used to gently nudge these users in the correct direction and to influence them to confirm to the standards of the community. This is because moderators can configure Automod to provide explanations for content removals to the users who posted them. Figure 3 shows an example of an explanation comment posted by Automod. Our participants told us that such explanations usually provide detailed descriptions of the subreddit rule the submitter has violated and the steps that can be taken to avoid such removals in the future. Participants noted that such explanations are often effective in helping users understand the social norms of the community. For example, PB_2 said:

> *"A lot of the time, it is just easier to follow the rules, and people that don't want to waste their time tend to conform to our requirements rather than go through the hassle."*

Over time, Automod has become an integral and indispensable part of the content regulation system on Reddit. It executes a large amount of menial work previously done by human moderators, and helps new users understand the norms of the subreddit. Next, we discuss how the syntactic configurations of Automod allow subreddits to automatically enforce some of their submission guidelines but not others.

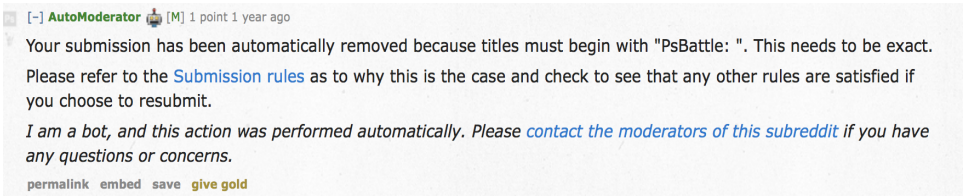

Fig. 3. A comment posted by Automod in response to its removal of a submission on r/photoshopbattles. Automod on r/photoshopbattles has been configured to automatically detect submissions that do not begin with 'PsBattle:', remove those submissions, and post this message as reason for removal.

5.3 Use of Automod to Enforce Community Guidelines

Community guidelines play an essential role in shaping the community and its moderation. They not only establish standards for how the users should behave on the subreddit, but they also set expectations for what users can count on from the moderation team. These guidelines (often called subreddit rules) directly affect how Automod is configured. That is, Automod rules are often created so that they can help ensure compliance with these guidelines.

> *"We have our rules in place and we make Automod conform to them as much as possible."*
> *– $ELIF_2$*

For example, the submission guideline #1 on r/photoshopbattles says, "All titles must begin with PsBattle:" [18] To ensure that users comply with this rule, moderators on r/photoshopbattles have configured an Automod rule that detects and removes all submissions that don't begin with 'PsBattle:' (see Figure 3). Here, it is worth noting that Automod not only removes an undesirable submission, it also automatically provides the poster the reason for removal, thereby serving an educational role [109]. Similarly, Automod helps ensure compliance with the r/explainlikeimfive guideline #10: "All Posts Must Begin With "ELI5."

> *"We already were reasonably strict in our rules, and a few of them were very easy to implement as Automod rules right out of the gate. The most obvious is that posts must start with "ELIF:" in our sub. We frequently removed posts without that prefix and asked them to repost it, and then suddenly Automod could do that perfectly and instantly with no work." – $ELIF_1$*

As another example, the r/politics guideline # 3: "Submissions must be from domains on the whitelist" requires that users should only post links from a specified list of websites that are considered appropriate for the subreddit[19]. The r/politics moderators ensure compliance with this guideline by configuring an Automod rule that checks whether each submission links to one of the domains on the whitelist. Another r/politics Automod rule guarantees compliance with the submission guideline #9: "Do not use "BREAKING" or ALL CAPS in titles" by checking for each submission whether its title contains all uppercase letters or if it contains the word "BREAKING."

While the above examples illustrate cases where Automod rules were created to enforce compliance with the existing guidelines, we also found a few instances where subreddit guidelines were specifically created to make content regulation easier to operationalize using Automod. For example, r/oddlysatisfying created a rule that requires submitters to describe the content shown in the image they submit in their post title. Enforcing this rule provides more information for Automod to detect and remove posts that are unsuitable for the r/oddlysatisfying community. Participant OS_1 explained how this rule helped ease the workload of moderators:

> *"The title rule helped make it easier to mod[erate] as it sped up the process of removing something that was NSFW[20] or had no purpose in the sub without having to look at or click every single submission."*

Here, we see how a community guideline was created to facilitate automated processing and alleviate the work of content moderators. This guideline, however, increases the input required of end-users by requiring them to provide a description of their image submissions. This is an example of how adopting automated tools to ease the work of moderators can create additional burdens on other stakeholders.

We found some variations in how different subreddits configure and use Automod. In Section 5.2.2, we showed the differences in the proportion of actions taken by Automod on various subreddits. Our interviews provided us additional insights. For example, OS_2 told us that r/oddlysatisying uses Automod only to detect "bad and threatening language." In contrast, PB_1 informed us that r/photoshopbattles configures Automod to help facilitate more sophisticated content curation, e.g., "Photoshops Only Mode" threads that only allow comments that are photoshopped versions of the image in original submission. In this case, Automod is configured to remove any comments that don't link to an image hosting site.

[18]This guideline helps distinguish r/photoshopbattles images from other pictures when they appear on the Reddit front page along with images from other subreddits.
[19]This whitelist is available at https://www.reddit.com/r/politics/wiki/whitelist.
[20]Not Safe For Work

> *"I set the thread to "Photoshops Only Mode" when the thread reaches 150 comments. That way, AutoModerator takes care of off-topic chains for me." – PB_1*

Not all policy guidelines are equally amenable to be enforced using Automod configurations. Moderators of every subreddit in our sample pointed out the grey areas in their guidelines that require subjective interpretations, and they consider such guidelines harder to implement using Automod. For example, $ELIF_3$ told us that r/explainlikeimfive moderators do not rely on Automod to enforce four (out of ten) subreddit guidelines[21] because they require a level of interpretation that are beyond the capabilities of Automod, e.g. guideline #4: "Explain for Laypeople" and guideline #5: "Explanations Must Be Objective." Instead, human moderators review all comments to ensure that these guidelines are followed.

A majority of our participants also noted that Automod rules are unable to consider context when making moderation decisions. For example, certain words or phrases can be used in multiple settings – their use may be acceptable to moderators in some contexts but not in others. $Space_1$ described how Automod removed a comment containing the term "shit" that he had to manually approve:

> *"Somebody talked about how they've "read this shit" in an explanation of their longstanding fascination with a topic. People that use that word tend to use it in mean-spirited or unserious comments, but this was an example where it's just for emphasis." – $Space_1$*

A few participants told us that their subreddits prohibit hate speech, but when moderating content posted to the subreddit, they sometimes find it difficult to determine whether a given comment should be considered hate speech or not. They try to attain a balance between allowing users to freely exchange ideas and prohibiting dialogue that make some users feel attacked or unsafe. In pursuit of this balance, however, participants find it challenging to use Automod to enforce guidelines prohibiting offensive behavior. For example, $ELIF_1$ shared:

> *"Our #1 rule is "Be nice," and we take that very seriously. That's definitely something that requires some interpretation. Someone saying "Way to be an idiot" is not nice. But is someone saying, "the way this works is X, people who think it works like Y are idiotic" not nice?"*

Knowing these limitations, moderators generally tend to configure Automod in a way that allows decisions that are less ethically ambiguous to be made automatically. Posts that are caught by Automod are removed as soon as they are posted, but many undesirable posts remain that are later removed by human moderators when they are reviewed. As a result, the use of Automod offloads some of the work of human moderators and frees them up for other tasks, e.g., making moderation decisions on posts that are harder to adjudicate.

> *"The goal of Automoderator is to get rid of the clear-cut things that you don't need a human to do. You can just look at the text and go, hey yeah, it's against the rules." – Pol_2*

> *"Automod does a lot of filtering of the worst stuff for us...It makes things easier and less stressful. We don't have to be trolling every thread for the worst stuff to get removed." – $Space_1$*

As the above quotes indicate, the use of Automod not just decreases the total amount of work that moderators need to do, it also reduces the emotional labor of moderators by minimizing their exposure to violent or offensive content.

Automod can be configured to either directly remove a post or triage it to a queue where human moderators can review that post. When Automod is used to flag a post for human review, the default decision (configurable in Automod) can be to either allow the concerned post or to automatically

[21] https://www.reddit.com/r/explainlikeimfive/wiki/detailed_rules

remove it until a human moderator gets to review that content. Therefore, the use of Automod provides some flexibility even when making difficult decisions. For example, OS_2 described a scenario in which Automod flags accounts for human review:

> *"Basically if a user's account is under X days old or has less than X karma, it will be automatically placed in the report queue to make sure it's not a spammer." – OS_2*

To sum up, Automod is well adept at enforcing some of the policy guidelines but not others. Subreddits still have to depend on human moderators to enforce guidelines that require subjective interpretation. This highlights the gap between what the syntactic instantiation of Automod rules can do and what the subreddit policies that are at a semantic level require. Next, we discuss the factors that influence how Automod is configured.

5.4 Mechanics of Automod Configuration

Given that adding or changing any rule of Automod can affect the moderation status of a large number of posts, we explored in our analysis how the decisions to edit Automod rules are made and who makes these edits.

5.4.1 Social Dynamics. Our interviews suggest that only a few moderators in each subreddit take on the responsibility of actively configuring Automod rules because it is difficult for others to understand how to configure it.

> *"All the active mods in ELIF can edit Automod, but few do because it's complicated, and our Automod config is pretty huge." – $ELIF_1$*

Participant PB_2 told us that he has handled most of the Automod coding of r/photoshopbattles for the past 4.5 years. Although that subreddit has a couple of other moderators who know how to program Automod, the rest of the moderation team considers PB_2 a single point of control for managing Automod code. PB_2 described that this arrangement allows the subreddit to identify and debug errors in Automod configurations more efficiently. He elaborated:

> *"I am a single point of control because [otherwise] if Automod starts going wrong then I have no idea what the issue is or how to fix it . . . Standards and good general practice make things easier to maintain but any time you get a group of people working on the same code without strict control, things get messy. It has just happened that I just handle it."*

In line with this, some participants from other subreddits told us that they are not informed of most of the changes in Automod configuration. They stressed, however, that they have the authority to reverse the decisions made by Automod if they deem them to be in error. For example, $Space_1$ stated:

> *"I am not always consulted as to what Automod will filter, but I can always override Automod."*

Six of our participants reported editing Automod rules. Participants with little knowledge or prior experience of editing Automod reported making only minor changes to existing Automod rules. For example, if there is an Automod rule that removes posts containing any word from a list of swear words, novice moderators feel comfortable adding another swear word to that list. This is because, as our participants pointed out, they have noticed Automod removing comments containing any swear word already on the list. They therefore recognize that their addition of a new swear word to the Automod rule will activate similar removals of all subsequent user comments containing that word. These moderators restrict themselves to making only minor changes, however, so as to ensure that their changes to Automod do not create errors in the functioning of Automod or result in unanticipated moderation actions. For example, OS_3 said:

> *"I don't get too involved with Automod. I know how to do the basic stuff, but not the coding side of things."*

In addition to restricting themselves to making only minor changes, novice moderators also usually inform the entire moderator team of the changes they make to Automod rules so that moderators with more experience of editing Automod can undo those changes, if required. Participants also told us that when they make changes to Automod rules, they often add documentation on the reasons for those changes as comments in the Automod code. Thus, moderators self-calibrate their skills at configuring Automod and practice care when making any changes to Automod rules.

As another example of moderators practicing care in their work, moderators often decide how to add or change any Automod rule based on their expectations of the number of posts that rule will affect. For example, Pol_1 told us that when he makes changes that are not expected to affect too many additional posts, he makes such configuration edits either by himself or by consulting with a few moderators. On the other hand, when he adds another domain to the white list of domains[22], this addition is discussed and voted upon by all the moderators before it is configured in Automod. This is because it is expected that many new posts may link to the domain in question and adding a rule to allow such posts may substantially affect the content available on the subreddit.

Participants from each subreddit in our sample noted that the moderator teams sometimes deliberate over the issues around the configuration of Automod rules in communication channels like Slack and Discord. However, such discussions are relatively rare. For example, $ELIF_1$ said:

> *"I don't think we've really had a discussion about Automod in a while. I see it like a mop for janitors. It's necessary, and you use it all the time, but there isn't really much to discuss about it."*

Half of our participants complained that Automod has a significant learning curve. Even moderators who understand regular expressions do not often use the advanced capabilities of Automod such as checking multiple conditions before triggering a rule. They either do not realize such configurations are possible or consider crafting such configurations too difficult. For example, Pol_1 said:

> *"It's not necessarily user-friendly... it almost entirely functions on regex, and its own little quirks and syntax to implement things so it can take some time for people to get decent at using it. There are a lot of mods whose eyes glaze over when having to work with it and [they] would rather do something else."*

As a result of the steep learning curve of configuring Automod, some moderators with little or no prior experience of editing Automod rely on other moderators for making the requisite changes. For example, $ELIF_2$ told us that moderators who do not know how to configure Automod often request other expert moderators to make changes in the configuration. $ELIF_2$ argued that complying with such requests added new responsibilities over moderators who take on the job of editing Automod without conferring any additional power or benefits to them. He further explained:

> *"Honestly, I feel like the mods who know more about bots get run over a bit. "Hey we need to add this, will @xyz or @yzz help please?" on Slack, or whatever."*

We found that moderators do not reveal the details of exactly how Automod works to their users. For example, the wiki page for Automod rules is not accessible to regular users by default. Our participants told us that although Reddit provides them the ability to make this wiki page public, they choose not to do so to avoid additional work and to ensure that bad actors don't game the Automod rules and post undesirable content that Automod cannot detect.

[22] As mentioned before, r/politics subreddit has set up a rule in Automod to allow only those submissions that link to one of the configured whitelist of domains.

> *"If you know what exactly is in the [Automod] code, then it is easily bypassed/exploited. Users with the inclination could figure out how to attack us maliciously or spam unhindered." – PB_2*

> *"It would be a massive pain in the ass to have it public, because it would require regular published updates, and we'd end up having to explain each modification to at least one of the few people complaining about censorship." – $ELIF_2$*

This necessary lack of transparency creates tensions between the moderators and the community. Many users are not aware of the presence of Automod, and posters whose comments are removed are usually not shown whether their comment was removed by Automod or a human moderator. Therefore, when users' comments get mistakenly removed by Automod, they often attribute such removals to human moderators and consider them unreasonable. Participant Pol_2 explained:

> *"It will be this big, incredible comment, and 98% of it will be, boy, this guy did his research, and he's doing good work, but it will get caught [by Automod] by something he says, and get removed. In those cases, it can turn into users thinking, 'Some moderators are filtering my speech — they don't like what I'm saying!' "*

By contrast, we observed that on many subreddits, when Automod removes a submission, the poster is notified that their submission was removed automatically. This is usually done through a comment to the removed submission authored by Automod (e.g., see Figure 3) or through automatically flairing the submission with a short removal reason [52]. Thus, moderators negotiate in distinct ways which aspects of the use of Automod they reveal to their users and how. This is an example of added responsibility and decision-making that moderators are obliged to perform when they adopt Automod.

5.4.2 Need for Careful Curation of Automod Rules. A majority of our participants felt that the utility of Automod largely depends on how its configuration rules are set up. Some moderators who are not too familiar with using regular expressions end up writing rules that are too broad and remove content that they didn't intend to remove.

> *"Sometimes, mods implement rules that accidentally remove too many things. In those cases, after a user has asked us to review a removal, I've gone back and refined the [Automod] rule to better capture what we're looking for. Regex (what the rules are made with) can be tricky." – $ELIF_1$*

Automod does not provide any feedback on the number of times a specific rule has been triggered. If the moderators don't pay attention to how their Automod rules affect the moderation on the subreddit by tracking the content that is being automatically removed, it may take them a while to realize the occurrence of unintended post removals. Seven participants from four different subreddits expressed their desire to get additional information about how Automod operates. They argued that analyzing statistical data about how different rules of Automod affect content regulation would allow them to fine-tune Automod configurations more efficiently.

> *"A poorly phrased regex bit can make something that looks like it shouldn't trigger on a post, trigger. But ... how do I know which one of the thirty five Automod rules did it? How do I know which part of the post made the trigger? ... I want to know which rules were invoked for which posts, how frequently, etc. - both in aggregate and on individual posts." – $ELIF_2$*

> *"If there was an easier way to see each removal reason and just a sampling of the comments that were removed for that removal reason, that would be pretty powerful. It would give you a better way to check your work." – Pol_2*

Moderators often rely on user reports to understand when an Automod configuration triggers too many unintentional side-effects. Although such user reports can allow moderators to correct their configurations, such mistakes create dissatisfaction among users.

> *"There was also a lot of users that were quite upset about it simply because they call it basically the censorship bot because it can just remove anything immediately with no ability for people to reason with it or convince them that it's the wrong decision." - Chad*

Three of our participants showed concern that many moderators focus on minimizing false negatives rather than false positives. In other words, moderators try to ensure that Automod is configured to catch as many undesirable posts and comments as possible, but they don't pay enough attention to whether Automod removes content that should not have been removed.

> *"I think, one of the things that bothered me before I was a moderator and complained about a lot was the false positives. Like, half the time, the Automoderator would have automatic comment removals if your comment had the word "homo" in it. But homo is not just a gay slur; it's also the genus of human beings – homo sapiens – so if you would write a comment using the word "homo sapiens" in it, your comment would be removed." – $Space_3$*

Here, we see that the moderators focus more on bad actors being punished, but ignore cases where good members are wronged. But the health of a community may be reliant on the latter just as much as on the former because undeserved punishment risks creating chilling effect and ultimately drives members away [55]. Additionally, as we will discuss below (Section 5.5), false positives also result in many user complaints, which consequently increases the moderators' workload of responding to those complaints.

In summary, since Automod only allows configuring keyword-based rules, the moderators have to make tradeoffs between (1) removing all posts that use the keyword in question including the posts that are acceptable or (2) not constituting a rule for that keyword and manually removing posts that are unacceptable, thereby increasing the work of human moderators. Thus, the use of Automod complicates the value-laden issues involved in content regulation.

5.5 Automod Creates New Tasks for Moderators

Although Automod certainly helps moderators to deal with the challenges of content regulation, its efficient deployment requires the moderator team to take on a set of new tasks in addition to configuration of Automod rules. We discuss these tasks in this section.

5.5.1 Regular Updating of Automod Rules. Participants told us that Automod rules on their subreddit have become more refined over time as the moderators continue tuning them to attain more accurate automated regulation decisions. Still, updating these rules is a continuous process - user content changes with the influx of new users and changes in cultural trends, and new requirements for automating specific moderation tasks are identified and configured for in Automod. Thus, Automod incorporates the historical evolution of the expectations for postings on each subreddit.

> *"95% of our Automod [code] changes are probably just adding, removing or refining items based on current events." – $ELIF_1$*

> *"Modifications to Automod code tend to come up on an as-need basis. If something slips by an existing rule, then the code for it is bolstered to cover that hole." – PB_2*

The wiki page for Automod rules records the history of each edit. This allows each subreddit to keep track of which moderators make what modifications to its Automod configuration. It also helps ensure that moderators who make any changes to Automod rules are accountable for their actions. It additionally allows quick reversion of any Automod rule changes, if needed.

Fig. 4. This comment shows the commenter's attempt to avoid detection by Automod. When $ELIF_1$ noticed this comment, he banned the commenter from r/explainlikeimfive subreddit. We have blotted usernames in the figure to preserve anonymity.

5.5.2 Preventing Users from Circumventing Automod. Our participants recognized that Reddit users can easily evade Automod rules by identifying the rule that triggers removals and use tactics like creative misspellings to bypass that rule. Therefore, they make efforts to guard against deliberate circumvention of Automod rules. As we discussed in Section 5.4.1, moderators do not provide users access to the wiki page where Automod rules are configured. Moderators of r/explainlikeimfive and r/photoshopbattles told us that they ban users who try to evade being caught by Automod. For example, $ELIF_1$ banned the user who posted the comment shown in Figure 4 because that user clearly attempted to bypass the Automod rule for removing any comments that are too short in length. In fact, some Automod rules on r/explainlikeimfive are configured to detect attempts to evade other rules and notify the moderators so that they can take appropriate actions against the suspected user.

5.5.3 Correcting False Positives. As we discussed in Section 5.4, moderators tend to focus on minimizing false negatives rather than false positives. As a result, Automod can often remove postings that do not violate community guidelines. Our participants reported that a bulk of their moderation mail contained complaints from new users about mistakes made by Automod. This creates additional work for the moderators by requiring them to respond to user complaints about the false positives of Automod's decisions.

> *"Valid complaints are usually about our bots getting a false positive for their issue as we have them set up pretty tight to make sure certain things don't slip by." – PB_2*

Thus, while using Automod allows moderators to offload a lot of their work and enact content regulation more efficiently, it also requires moderators to develop new skills like configuring Automod rules and conduct additional activities like defending against deliberate avoidance of Automod filters and correcting false positives. Since these new tasks are not trivial, the use of Automod creates new challenges of training and coordination among moderators.

In summary, although Reddit moderators can regulate their communities without using any automated tools, the use of these tools makes their work more convenient and manageable. The combination of natural human abilities of moderators with the capacities of external components like Automod and Reddit toolbox forms a system that performs the existing function of content regulation more efficiently [59]. In their research on the use of automated tools in Wikipedia

governance, Geiger and Ribes noted that "the delegation of certain tasks to these tools makes certain pathways of action easier for vandal fighters and others harder" [36]. Similarly, we found that using automated tools on Reddit changes the underlying activity of content regulation and raises new challenges (such as preventing users from evading Automod rules) that moderators need to grapple with.

6 DISCUSSION

Our research extends prior work on content moderation by drawing attention to the automated regulation tools that moderators use. We describe the sociotechnical practices that shape the use of these tools. We also highlight how these tools help workers maintain their communities. Our analytic approach has allowed us to identify the limitations of current automated systems and recognize the important design challenges that exist in attaining successful moderation. In this section, we describe these challenges and limitations. We also propose solutions that may help address them. Furthermore, we discuss what other communities that incorporate automated tools in their regulation systems may learn from our research.

6.1 Facilitating Development and Sharing of Automated Regulation Tools

Centralized moderation tools and mechanisms are often developed using universalist design principles and practices that assume that the 'default' imagined users belong to the dominant social groups [25]. Yet, these official moderation tools may not be able to satisfy the requirements of all communities [55]. Moderators are well-poised to identify these social-technical gaps [1] because they work closely with their communities and can recognize the specific needs of their users that official moderation tools do not satisfy. Therefore, mechanisms that allow these moderators to develop and deploy regulation tools that meet the unique requirements of their communities can substantially improve content regulation.

One such mechanism is to provide moderators an API access to the community data. Our findings suggest that the open and flexible API provided by Reddit platform has encouraged the development of a wide variety of automated tools. Volunteer users frequently create moderation bots that tailor to their community and improve its regulation. Similar to this, Jhaver et al. found that Twitter blocklists, a third-party moderation tool developed by volunteer users on Twitter helped enhance the experiences of many marginalized users by allowing them to curate content in ways that were not possible through centralized moderation mechanisms officially offered by Twitter [55]. Therefore, we recommend that more platforms should consider providing API access that volunteer developers can use to build and deploy automated regulation bots that meet the specific needs of their communities. Opening up content regulation to third-party developers should also encourage implementation and testing of creative new ideas for community management.

We found that Reddit moderators spend considerable time and efforts developing bots to improve content regulation for their own subreddit using automated mechanisms. But bots that are valuable for one community can also bring immense value to regulation of other communities. For example, we saw that although Automod was initially developed for r/gaming subreddit, it eventually became an indispensable part of the Reddit regulation system. Still, as we discussed in our findings, there is no central repository of all the automated tools that moderators can directly use. Moderators only come to know about such tools through their contacts with moderators in other subreddits. This results in duplicate effort on the part of bot developers. To avoid such duplication, platforms like Reddit should encourage volunteer developers to build tools that can be quickly adapted to enact regulation in other similar settings. Platforms may also promote sharing of such tools on a centralized repository so that other moderators can directly access them and adapt them for their own communities.

6.2 The Need for Performance Data

Our findings show that there is lack of accessible data on how well the automated parts of regulation systems on Reddit work. Currently, Automod does not provide any visibility into the number of times each rule has been triggered. This is problematic because rules added to Automod sometimes have unintended effects. We found that because of the absence of easy visibility into Automod behavior, moderators often have to rely on user reports to identify the occurrence of unintended post removals. This highlights the importance of flagging mechanisms [26] and contributions of users to content regulation [66] even in systems that rely on automated tools. Yet, as our findings show, mistakes made by Automod can frustrate the users affected.

To facilitate quick discovery of Automod mistakes, we recommend that designers build audit tools that provide moderators visibility into the history of how each Automod rule affects the moderation on the subreddit. Such visibility would allow moderators to edit Automod rules if required and control its actions more closely. Audit tools could also be enhanced to show moderators the potential consequences of creating a new rule by simulating application of that rule on already existing data in sandpit type environments. If moderators are able to visualize the type of comments that would be removed by creation of a new rule, they would be better positioned to avoid crafting broad rules that result in many false positives.

We expect that the same principle also applies to automated regulation on systems other than Reddit. It is vital for human moderators to understand how different configurations of automated regulation systems affect the curation of their sites. As our findings show, moderators want the ability to quickly locate the settings that result in undesirable regulation decisions and fix them. Therefore, automated systems should be designed so that moderators have detailed visibility into how automation affects content curation. Moderators should also be able to tune the configurations of such systems at a granular level and maintain control over how these systems work.

While the provision of performance data, as discussed above, is important to evaluate automated moderation, it should be noted that systematic evaluation of moderation records, in general, is a non-trivial endeavor. A thorough evaluation would require sampling and reviewing of posts that have been allowed to be kept up as well as posts that have been removed - either by automated tools or by a human moderator. For each moderation action, different moderators may have different views of whether that action is appropriate - such conflicts and disagreements are part of the political process of enacting moderation. In addition, it is possible that a majority of users may disagree with the moderators' decisions. Thus, post-hoc evaluation of content moderation records as a social practice has many critical concerns that need careful examination, and is a ripe area for future research.

We also found that only a few technically adept moderators can configure Automod, and many subreddits are unable to tap into the full potential of Automod. This is similar to Geiger and Ribes' finding on automated regulation in Wikipedia that while many "workarounds are possible, they require a greater effort and a certain technical savvy on the part of their users" [36]. Therefore, we recommend that automated systems be designed in such a way that moderators can easily understand and configure their settings. This would allow more moderators to engage with automated systems, and facilitate conditions where a larger share of moderators can influence content curation using automated tools.

6.3 Human versus Automated Moderation Systems

Our analysis shows that identifying tasks that should be automated and configuring tools to perform those tasks is crucial for Reddit moderators' ability to maintain their communities, especially as the communities grow large. This finding is consistent with Epstein and Leshed's observation

that automation of maintenance tasks like detecting and addressing incivility have been critical to scaling up the RegulationRoom[23] deliberation environment [31]. While automation is crucial to support the growth of online communities, accurate automated detection is not an easy task. It is arguably impossible to make perfect automated moderation systems because their judgments need to account for the context, complexity of language and emerging forms of obscenity and harassment, and they exist in adversarial settings [76] where they are vulnerable to exploitation by bad actors.

Automating moderation not only facilitates scalability, it also enables consistency in moderation decisions. In a recent Data & Society report, Robyn Caplan noted that many companies tend to formalize their logic in order to address regulation concerns more consistently [13]. These companies transform content moderation standards into hard-coded training materials for new workers as well as automated flagging systems. This is in line with how moderation criteria on Reddit are specified as fixed regular expressions in Automod rules. Although hard-coding moderation criteria facilitates scalability and consistency of moderation systems, such transformation of content moderation values can end up being insensitive to the individual differences of content, for example, when distinguishing hate speech from newsworthiness [13]. These failures to address context issues can have serious consequences, e.g., persistence of misinformation campaigns on Facebook or WhatsApp that arguably contributed to violence in Myanmar [101].

More advanced automated systems that use machine learning models, especially those based on deep-learning frameworks, currently cannot provide specific reasons for why each removed comment or post was removed. Advances in human-understandable machine learning may help address this problem in the future [64]. Currently, our findings show that moderators adopt Automod because they can directly control how it works by editing its configuration. They can understand the mistakes made by Automod by observing the keywords that triggered those mistakes and explain such mistakes to placate dissatisfied users. Prior research has also shown how retaining control over content regulation is important to the moderators [29].

Therefore, we caution researchers and designers that although AI moderation systems are invaluable for managing many moderation tasks and reducing workload [99], deploying such systems without keeping any humans in the loop may disrupt the transparency and fairness in content moderation that so many users and moderators value. This is in line with speculations made by other researchers that machine-learning driven moderation approaches are inherently risky because they may "drive users away because of unclear or inconsistent standards for appropriate behavior" [96]. Additionally, we found that only a small number of moderators in each subreddit configure Automod because others do not have the technical expertise to make such configurations. The use of more complex machine learning tools can further raise the bar for users who can moderate online communities while also disproportionately increasing the workload of moderators who can work with those tools. Thus, platforms should consider that they may lose valuable moderators when moving to systems that heavily rely on machine learning tools.

Designers and moderators must recognize that the use of automated regulation systems fundamentally changes the work of moderators. For example, when subreddits use Automod, moderators' work becomes constrained to adjudicate only those postings that are not caught and removed by Automod. Moreover, it creates additional tasks that require technical expertise such as regular updating of Automod rules and preventing users from circumventing Automod. We confirm the finding of Seering et al. [97] that Automod sometimes adds to the work of moderators because they have to manually approve content mistakenly removed by Automod. Therefore, when moderators incorporate new automated mechanisms in their content regulation systems, they should anticipate

[23] http://regulationroom.org

new tasks and prepare to execute and train for those tasks. More generally, while moderators and new automated systems may co-evolve and adapt to each other, it is still important to consider the resulting social-technical gaps as CSCW problems and make efforts to "round off the edges" of coevolution [1, 88].

We found that Reddit moderators show some aspects of the work of Automod to their users but not others. These decisions are important in order to retain the trust of the users while at the same time ensuring that bad actors do not game the system and bypass Automod rules. Concerns about users evading automated moderation systems are not limited to Reddit — prior research has shown how bad actors on Instagram and Tumblr circumvent platforms' efforts to moderate problematic hashtags by devising innovative ways to promulgate controversial content [16, 37]. Therefore, when incorporating automated regulation systems, moderators should be prepared to make critical decisions about which parts of the automated system to show and which to hide from their users.

Our findings also bring attention to the tradeoffs between reducing the work of human moderators and not automatically removing posts that may potentially be valuable despite having suspicious characteristics. On one hand, using Automod reduces the amount of work that Reddit moderators need to do and protects them from the emotional labor of scrolling through the worst of the internet's garbage. On the other hand, it is all too easy for moderators to configure rules that are too broad in Automod. Although such a configuration catches and removes many potentially unacceptable posts and reduces the dependency on human moderators, it results in many false positives that may alienate users. We also found that human moderators are needed to frequently update Automod rules so that Automod can account for the fluidity of culture and adaptability of violators seeking to avoid detection.

6.3.1 Improving Mixed-Initiative Systems. Given the deficiencies of automated tools and the importance of careful human administering of these tools, we propose that instead of developing fully automated systems, researchers and designers should make efforts to improve the current state of mixed-initiative regulation systems where humans work alongside automated systems. Since automated tools are likely to perform worse than humans on difficult cases where understanding the nuances and context is crucial, perhaps the most significant consideration is determining when automated tools should remove potentially unacceptable material by themselves and when they should flag it to be reviewed by human moderators. It is critical for these tools to attain this balance to ensure that unintended post removals are avoided and at the same time, the workload of human moderators is substantially reduced. We echo calls by previous studies for building systems that ensure that the interactions between automation and human activities foster robust communities that function well at scale [31, 35, 97].

Another promising direction to explore would be to build systems that adapt hybrid crowd-machine learning classifiers like Flock (developed by Cheng and Bernstein [21]) for the purpose of content regulation. Such systems would require a dataset of comments that have been thoroughly moderated and labeled as 'approved' or 'removed'. To begin with, such a system would guide human moderators to nominate effective features for distinguishing approved posts from removed posts. This would be followed by the use of machine learning techniques that weigh these features and produce models that have good accuracy as well as recall and that use human-understandable features. Our findings indicate that moderators would appreciate the ability to understand outputs based on such features. As Cheng and Bernstein suggest [21], the performance of these models could be further improved by identifying spaces where misclassifications occur. Moderators could be asked to nominate additional features that may be informative in improving performance in

those spaces. Researchers and practitioners could build, deploy and test such systems on social media platforms, and compare their performance with existing regulation mechanisms.

6.4 Limitations and Future Work

This study has some limitations. Our results are from interviews with a small sample of Reddit moderators. We note a self-selection bias: we only spoke with Reddit moderators who were willing to talk to us. Our sample is diversified in that our participants host a variety of subreddits, come from various geographic areas, and have different occupations. Participants not only moderate the five subreddits that we sampled from, but also a number of other subreddits (Table 2). Nevertheless, our sample was all male. We suspect this is because Reddit moderators are disproportionately male; still, to our knowledge, no comprehensive data on the demographics of Reddit moderators exist. Prior research has shown that gender shapes individuals' conceptions of offense in online posts [7]. Besides, many scholars have studied the issues of gender equity in online forums and their effects on democratic discourse [47, 72, 87, 110]. Therefore, future work on the demographics of Reddit moderators and especially how gender affects moderation practices would be a useful contribution.

As we progressed through our interviews, we began to hear the same themes again and again. Our final few interviews generated limited new insights, suggesting our data reached empirical saturation. This supports the validity of our results. Additionally, we note that social desirability bias might have affected what our participants were willing to share with us.

One rich direction for future work is to evaluate the performance of Automod and characterize its false positives and false negatives. Moderators could be asked to code whether they would allow or remove a sample of postings on their subreddit, and these codes could be compared with the actual outcomes of Automod processing of those postings. This could provide useful insights into the net workload saved because of the use of Automod and the amount of new workload generated because of the false positives of Automod's decisions. Additionally, Automod's performance could be compared with the results of machine learning models trained on previously moderated data from the subreddit. Analyzing posts on which moderators disagree or find it difficult to take a decision could also provide valuable insights about moderation.

Finally, this paper presents the point of view of moderators. In future work, it would be beneficial to study the perspectives of participants whose comments and posts may or may not be deleted by Automod. More generally, analyzing the effects of adopting automated moderation tools on the design of posting guidelines and the demands on end-users on different platforms would provide valuable insights. It would also be useful to investigate when platforms' interests align with and differ from those of volunteer moderators.

7 CONCLUSION

This paper presents a qualitative inquiry of the content regulation ecosystem on Reddit, one of the most popular social media platforms. Our findings show that content regulation on Reddit is a socially distributed endeavor in which individual moderators coordinate with one another as well as with automated systems. We discuss the benefits of automated tools and identify areas for improvement in the development and sharing of these tools.

We see this research speaking to multiple stakeholders: creators of platforms, designers and researchers interested in automated or machine learning-based content regulation, scholars of platform governance, and content moderators. In conclusion, we highlight the following takeaways from our work:

7.1 For creators of new and existing platforms

As user traffic increases, moderation systems have to process content at massive levels of scale. This makes it necessary for platforms to use automated tools. Our findings suggest, however, that the use of automated tools has direct and secondary effects on multiple stakeholders and their activities — from how moderators coordinate among one another and create community guidelines to how users are required to craft their posts. We therefore recommend that platforms carefully reflect on the anticipated ripple effects over different stakeholders when determining which automated tools they deploy in content regulation systems.

Usually, how content regulation occurs on social media platforms remains a trade secret and is not revealed publicly. In this research, we provide details of how Reddit moderators distribute the work of content regulation between human workers and automated tools. Our comprehensive description of Reddit regulation provides an important reference point for how human-machine mixed initiative regulation systems can be designed and deployed on other platforms.

7.2 For designers and researchers interested in automated content regulation

We highlight the concerns that designers of machine learning based content regulation should take into account when creating new tools. We found that although Automod relies on syntactic rules instead of advanced machine learning techniques, moderators value Automod because it provides them a great level of control and understanding of the actions taken by Automod. Our findings reveal that moderators who do not understand how automated tools work may not be able to contribute as much after these tools are adopted. This can, in turn, affect the dynamics of relationships among the moderator team. This highlights the significance of creating tools whose configurations are easily understood by the moderators, and designing tutorials that assist this understanding.

Furthermore, it is important to explore how the use of automated tools shapes the explainability of moderation decisions and the perceptions of affected users [24, 55, 56]. We also hope to see studies that investigate how the use of automated mechanisms differs between Reddit and platforms that rely on commercial content moderation firms.

7.3 For scholars of platform governance

In recent years, researchers have begun asking questions about the democratic accountability of platform companies and their role in the realization of important public values like freedom of expression, transparency, diversity, and socio-economic equality [39, 43, 46, 52, 103, 109]. Our findings contribute to this conversation by showing that an increased reliance on automated moderation tools can contribute to situations where content moderation may seem unfair. Since automated tools can't always consider the context of a post, they may consistently censor individuals with certain viewpoints, or they may influence the discursive norms in unforeseen ways and increase online polarization [7]. On the other hand, these tools may catch and remove posts that are problematic only at the surface level but allow proliferation of bigoted viewpoints that are subtle and avoid automatic detection at deeper levels of meaning. Exactly how the adoption of various types of automated moderation tools affects different user groups is a subject that scholars of platform governance must examine so that they can articulate strategies that may address the problems of tool bias.

Automated moderation tools not only exacerbate biases but they also operate simply by reacting to problems, not by dealing with their root causes. Such an approach simply hides problematic behaviors such as sexism and racism instead of interfacing with offenders in meaningful ways [50]. This may merely push the offensive users to other platforms where their bigoted views are more

welcomed. In this way, current automated tools miss out on the opportunities to examine the social and psychological factors that lead to hateful discourses. Instead, we call for researchers to find ways, which may go well beyond simply a deployment of automated tools, to change offensive or uninformed users' perspectives on socially relevant issues [53, 57, 95] and help shift norms in positive ways.

7.4 For content moderators

Our findings point out the new challenges that moderators can expect to grapple with as they adopt new automated tools in their work. Content moderation is *hard*, and even without the use of automated mechanisms, when moderators update their policies, they sometimes have to revisit previous positions [39]. However, delegating content regulation to automated tools increases the distance between how community guidelines are conceptualized by moderators and how they are enforced in practice by automated tools. This distance raises the possibility that moderation systems will make mistakes. Consequently, moderators can expect to take on the additional tasks of correcting the false positives of these tools.

The use of automated tools not only adds to the moderators' work by requiring them to reverse mistakes made by automated moderation, it may also affect the relationship between the users and the moderators [51, 100] because of the higher occurrences of mistakes. Using these tools also results in increased user complaints that moderators have to respond to, which further adds to the moderators' work. Besides, moderators may not have control over what is disclosed about the operation of these tools to the ends-users. Although the extent of such disclosures may instead be determined by designers or site administrators, it may still shape end-users' perceptions of moderators' work because end-users may consider moderators responsible for the choices of transparency in moderation decisions. Moderators may also need to develop technical expertise to use advanced machine-learning based tools efficiently. In sum, the use of automated tools changes the work required of moderators and their relationships with end-users in important ways. As community managers inevitably move towards adopting more automated tools for content regulation, efforts to prepare moderators for such changes will be vital.

ACKNOWLEDGMENTS

We would like to thank Jialun "Aaron" Jiang, Michaelanne Dye, Sucheta Ghoshal, Eshwar Chandrasekharan and Benjamin Sugar for their valuable inputs that improved this work. We would like to acknowledge Chad Birch for creating Reddit Automoderator and thank all our interviewees for taking the time to share their experiences with us. We also appreciate the Associate Editor and the reviewers of this article for their constructive feedback and encouragement. Jhaver and Gilbert were supported by the National Science Foundation under grant IIS-1553376.

REFERENCES

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.
[2] Alexa. 2018. reddit.com Traffic Statistics. https://www.alexa.com/siteinfo/reddit.com
[3] Automoderator. 2018. Automoderator - reddit.com. https://www.reddit.com/wiki/automoderator
[4] Oren Ben-Kiki, Clark Evans, and Brian Ingerson. 2005. YAML Ain't Markup Language. http://yaml.org/spec/1.2/spec.html
[5] Zane L Berge and Mauri P Collins. 2000. Perceptions of e-moderators about their roles and functions in moderating electronic mailing lists. *Distance Education* 21, 1 (2000), 81–100.
[6] BigQuery. 2018. Google BigQuery. https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments
[7] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International Conference on Social Informatics*. Springer, 405–415.

[8] Lindsay Blackwell, Jill P Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI* 1, CSCW (2017), 24–1.

[9] Amy Bruckman. 2006. Teaching students to study online communities ethically. *Journal of Information Ethics* (2006), 82.

[10] Amy Bruckman, Kurt Luther, and Casey Fiesler. 2015. When Should We Use Real Names in Published Accounts of Internet Research? In *Digital Research Confidential: The Secrets of Studying Behavior Online*, Eszter Hargittai and Christian Sandvig (Eds.).

[11] Catherine Buni. 2016. The secret rules of the internet: The murky history of moderation, and how it's shaping the future of free speech. https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech

[12] Catherine Buni. 2019. The Trauma Floor: The secret lives of Facebook moderators in America. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[13] Robyn Caplan. 2018. Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches. https://datasociety.net/output/content-or-context-moderation/

[14] captainmeta4 (Submitter). 2016. What is /u/BotBust? : BotBust. https://www.reddit.com/r/BotBust/comments/5092dg/what{_}is{_}ubotbust/

[15] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM Press, New York, New York, USA, 3213–3226. https://doi.org/10.1145/3025453.3025985

[16] 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*. ACM, 1201–1213.

[17] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.

[18] 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 SIGCHI conference on Human factors in computing systems*.

[19] Kathy Charmaz. 2006. Coding in Grounded Theory Practice. *Constructing Grounded Theory* (2006), 42–70.

[20] Adrian Chen. 2014. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed | WIRED. https://www.wired.com/2014/10/content-moderation/

[21] Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid Crowd-Machine Learning Classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 600–611. https://doi.org/10.1145/2675133.2675214

[22] Danielle Keats Citron. 2014. *Hate crimes in cyberspace*. Harvard University Press.

[23] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing revenge porn. *Wake Forest L. Rev.* 49 (2014), 345.

[24] Maxime Clément and Matthieu J Guitton. 2015. Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior* 50 (2015), 66–75.

[25] Sasha Costanza-Chock. 2018. Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice. (2018). https://ssrn.com/abstract=3189696

[26] 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.

[27] Deimorz (Submitter). 2012. AutoModerator - a bot for automating straightforward reddit moderation tasks and improving upon the existing spam-filter : TheoryOfReddit. https://www.reddit.com/r/TheoryOfReddit/comments/onl2u/automoderator_a_bot_for_automating/

[28] Jean-Yves Delort, Bavani Arunasalam, and Cecile Paris. 2011. Automatic moderation of online discussion sites. *International Journal of Electronic Commerce* 15, 3 (2011), 9–30.

[29] Nicholas Diakopoulos and Mor Naaman. 2011. Towards Quality Discourse in Online News Comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/1958824.1958844

[30] 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*. ACM, 142.

[31] Dmitry Epstein and Gilly Leshed. 2016. The magic sauce: Practices of facilitation in online policy deliberation. *Journal of Public Deliberation* 12, 1 (2016), 4.

[32] Casey Fiesler, Jialun Aaron Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*. 72–81.
[33] National Science Foundation. 2019. Future of Work at the Human-Technology Frontier: Core Research (FW-HTF). https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505620
[34] Sandra E. Garcia. 2018. Ex-Content Moderator Sues Facebook, Saying Violent Images Caused Her PTSD. https://www.nytimes.com/2018/09/25/technology/facebook-moderator-job-ptsd-lawsuit.html
[35] R. Stuart Geiger and Aaron Halfaker. 2013. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?. In *Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13)*. ACM, New York, NY, USA, Article 6, 6 pages. https://doi.org/10.1145/2491055.2491061
[36] 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
[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] Tarleton Gillespie. 2017. Governance of and by platforms. *Sage handbook of social media* (2017).
[39] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[40] Tarleton Gillespie. 2018. The Logan Paul YouTube controversy and what we should expect from internet platforms. https://www.vox.com/the-big-idea/2018/1/12/16881046/logan-paul-youtube-controversy-internet-companies
[41] April Glaser. 2018. Want a Terrible Job? Facebook and Google May Be Hiring. https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html
[42] Kirsten Gollatz, Felix Beer, and Christian Katzenbach. 2018. The turn to artificial intelligence in governing communication online. (2018).
[43] Robert Gorwa. 2019. What is platform governance? *Information, Communication & Society* 22, 6 (2019), 854–871.
[44] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[45] Hugo Lewi Hammer. 2016. Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*. Springer, 164–173.
[46] Natali Helberger, Jo Pierson, and Thomas Poell. 2018. Governing online platforms: From contested to cooperative responsibility. *The information society* 34, 1 (2018), 1–14.
[47] Susan C Herring. 2000. Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal* 18, 1 (2000), 0.
[48] Arlie Russell Hochschild. 1983. *The managed heart: Commercialization of human feeling*. Berkeley: University of California Press.
[49] Maya Holikatti, Shagun Jhaver, and Neha Kumar. 2019. Learning to Airbnb by Engaging in Online Communities of Practice. *Under Review at Proceedings of the ACM on Human-Computer Interaction* (2019).
[50] Matthew W Hughey and Jessie Daniels. 2013. Racist comments at online news sites: a methodological dilemma for discourse analysis. *Media, Culture & Society* 35, 3 (2013), 332–347.
[51] Shagun Jhaver, Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": User Reactions to Content Removals on Reddit. *Under Review at Proceedings of the ACM on Human-Computer Interaction* (2019).
[52] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Under Review at Proceedings of the ACM on Human-Computer Interaction* (2019).
[53] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (2018). http://firstmonday.org/ojs/index.php/fm/article/view/8232
[54] Shagun Jhaver, Justin Cranshaw, and Scott Counts. 2019. Measuring Professional Skill Development in U.S. Cities Using Internet Search Queries. In *Thirteenth International AAAI Conference on Web and Social Media*.
[55] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 12.
[56] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic anxiety and coping strategies of Airbnb hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 421.
[57] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for Civil Conversations: Lessons Learned from ChangeMyView*. Technical Report. Georgia Institute of Technology.
[58] Ling Jiang and Eui-Hong Han. 2019. ModBot: Automatic Comments Moderation. In *Computation+ Journalism Symposium*.

[59] Victor Kaptelinin. 1996. Activity theory: Implications for human-computer interaction. *Context and consciousness: Activity theory and human-computer interaction* 1 (1996), 103–116.

[60] Aphra Kerr and John D Kelleher. 2015. The recruitment of passion and community in the service of capital: Community managers in the digital games industry. *Critical Studies in Media Communication* 32, 3 (2015), 177–192.

[61] 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

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

[63] Kate Klonick. 2017. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131 (mar 2017). https://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=2937985

[64] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press, New York, New York, USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[65] Cliff Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the Reader: Filtering Comments on Slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1253–1262. https://doi.org/10.1145/1240624.1240815

[66] 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

[67] 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

[68] 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 *Twelfth International AAAI Conference on Web and Social Media*.

[69] Claudia Lo. 2018. *When all you have is a banhammer: the social and communicative work of volunteer moderators*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[70] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could You Define That in Bot Terms"?: Requesting, Creating and Using Bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3488–3500. https://doi.org/10.1145/3025453.3025830

[71] Alexis Madrigal. 2018. Inside Facebook's Fast-Growing Content-Moderation Effort. https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/

[72] Fiona Martin. 2015. Getting my two cents worth in: Access, interaction, participation and social inclusion in online news commenting. https://isojjournal.wordpress.com/2015/04/15/getting-my-two-cents-worth-in-access-interaction-participation-and-social-inclusion-in-online-news-commenting/

[73] Nathan J Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

[74] Nathan J. Matias. 2016. Posting Rules in Online Discussions Prevents Problems & Increases Participation. https://civilservant.io/moderation_experiment_r_science_rule_posting.html

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

[76] Patrick McDaniel, Nicolas Papernot, and Z. Berkay Celik. 2016. Machine Learning in Adversarial Settings. *IEEE Security & Privacy* 14, 3 (may 2016), 68–72. https://doi.org/10.1109/MSP.2016.51

[77] Aiden McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. *ICIS 2016 Proceedings* (dec 2016). http://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23

[78] Steven Melendez. 2015. Here's How 20,000 Reddit Volunteers Fight Trolls, Spammers, And Played-Out Memes. https://www.fastcompany.com/3048406/heres-how-20000-reddit-volunteers-fight-trolls-spammers-and-played-out-memes

[79] Sharan B Merriam. 2002. Introduction to Qualitative Research. *Qualitative research in practice: Examples for discussion and analysis* 1 (2002).

[80] Elise Moreau. 2017. What Exactly Is a Reddit AMA? https://www.lifewire.com/what-exactly-is-a-reddit-ama-3485985

[81] Kevin Morris. 2015. Reddit moderation being taken over by bots-and that's a good thing. https://www.dailydot.com/news/reddit-automoderator-bots/

[82] Enid Mumford. 2000. A socio-technical approach to systems design. *Requirements Engineering* 5, 2 (2000), 125–133.
[83] Erica Ong. 2018. Is Machine Learning the Future of Content Moderation? https://insights.conduent.com/conduent-blog/is-machine-learning-the-future-of-content-moderation
[84] M Pilar Opazo. 2010. Revitalizing the Concept of Sociotechnical Systems in Social Studies of Technology. (2010).
[85] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1114–1125.
[86] Michael Quinn Patton. 1990. *Qualitative evaluation and research methods*. SAGE Publications, Inc.
[87] Emma Pierson. 2015. Outnumbered but well-spoken: Female commenters in the New York Times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1201–1213.
[88] Neil Postman. 1992. *Technopoly*. New York: Vintage.
[89] Kim Renfro. 2016. For whom the troll trolls: A day in the life of a Reddit moderator. https://www.businessinsider.com/what-is-a-reddit-moderator-2016-1#crocker-has-
[90] Sarah T. Roberts. 2014. *Behind the screen: the hidden digital labor of commercial content moderation*. Ph.D. Dissertation. University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/handle/2142/50401
[91] Sarah T. Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications* (jan 2016). https://ir.lib.uwo.ca/commpub/12
[92] Sarah T. Roberts. 2019. *Behind The Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
[93] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities.. In *Proceedings of the 11th ACM Conference on Web Science*.
[94] Marcos Rodrigues Saude, Marcelo de Medeiros Soares, Henrique Gomes Basoni, Patrick Marques Ciarelli, and Elias Oliveira. 2014. A strategy for automatic moderation of a large data set of users comments. In *2014 XL Latin American Computing Conference (CLEI)*. IEEE, 1–7.
[95] 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*. ACM, 606.
[96] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277
[97] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019).
[98] Monika Singh, Divya Bansal, and Sanjeev Sofat. 2016. Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining* 6, 1 (2016), 41.
[99] Devin Soni and Vivek K. Singh. 2018. See No Evil, Hear No Evil: Audio-Visual-Textual Cyberbullying Detection. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 164 (Nov. 2018), 26 pages. https://doi.org/10.1145/3274433
[100] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* (2019).
[101] Steve Stecklow. 2018. Why Facebook is losing the war on hate speech in Myanmar. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/
[102] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research*. Sage publications.
[103] 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), 18.
[104] Courtnie Swearingen and Brian Lynch. 2018. We're Reddit Mods, and This Is How We Handle Hate Speech. https://www.wired.com/2015/08/reddit-mods-handle-hate-speech/
[105] Steven J Taylor, Robert Bogdan, and Marjorie DeVault. 2015. Participant Observation: In the Field. In *Introduction to qualitative research methods: A guidebook and resource*. John Wiley & Sons, Chapter 3.
[106] TL Taylor. 2018. Regulating the networked broadcasting frontier. In *Watch me play: Twitch and the rise of game live streaming*. Princeton University Press, Chapter 5.
[107] Ken Thompson. 1968. Programming Techniques: Regular expression search algorithm. *Commun. ACM* 11, 6 (jun 1968), 419–422. https://doi.org/10.1145/363347.363387
[108] Adriano Veloso, Wagner Meira Jr, Tiago Alves Macambira, Dorgival O Guedes, and Hélio Marcos Paz de Almeida. 2007. Automatic Moderation of Comments in a Large On-line Journalistic Environment.. In *ICWSM*.
[109] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* (2018).
[110] Vanessa L Wilburn. 1994. *Gender and anonymity in computer-mediated communication: participation, flaming, deindividuation*. Ph.D. Dissertation. University of Florida.

[111] 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*. ACM, 160.

[112] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399. https://doi.org/10.1145/3038912.3052591

Received January 2019; revised March 2019; accepted May 2019