

# Platform Governance with Algorithm-Based Content Moderation: An Empirical Study on Reddit

Qinglai He,<sup>a,\*</sup> Yili Hong,<sup>b</sup> T. S. Raghu<sup>c</sup>

<sup>a</sup> Wisconsin School of Business, University of Wisconsin–Madison, Madison, Wisconsin 53706; <sup>b</sup> Miami Herbert Business School, University of Miami, Coral Gables, Florida 33146; <sup>c</sup> W.P. Carey School of Business, Arizona State University, Tempe, Arizona 85287

\*Corresponding author

Contact: [qinglai.he@wisc.edu](mailto:qinglai.he@wisc.edu),  <https://orcid.org/0000-0002-4967-6453> (QH); [khong@miami.edu](mailto:khong@miami.edu),  <https://orcid.org/0000-0002-0577-7877> (YH); [Raghu.Santanam@asu.edu](mailto:Raghu.Santanam@asu.edu),  <https://orcid.org/0000-0003-1071-6339> (TSR)

Received: January 19, 2021

Revised: June 21, 2022; August 13, 2023

Accepted: November 25, 2023

Published Online in Articles in Advance:  
August 21, 2024

<https://doi.org/10.1287/isre.2021.0036>

Copyright: © 2024 INFORMS

**Abstract.** With increasing volumes of participation in social media and online communities, content moderation has become an integral component of platform governance. Volunteer (human) moderators have thus far been the essential workforce for content moderation. Because volunteer-based content moderation faces challenges in achieving scalable, desirable, and sustainable moderation, many online platforms have recently started to adopt algorithm-based content moderation tools (bots). When bots are introduced into platform governance, it is unclear how volunteer moderators react in terms of their community-policing and -nurturing efforts. To understand the impacts of these increasingly popular bot moderators, we conduct an empirical study with data collected from 156 communities (subreddits) on Reddit. Based on a series of econometric analyses, we find that bots augment volunteer moderators by stimulating them to moderate a larger quantity of posts, and such effects are pronounced in larger communities. Specifically, volunteer moderators perform 20.9% more community policing, particularly over subjective rules. Moreover, in communities with larger sizes, volunteers also exert increased efforts in offering more explanations and suggestions after their community adopted bots. Notably, increases in activities are primarily driven by the increased need for nurturing efforts to accompany growth in subjective policing. Moreover, introducing bots to content moderation also improves the retention of volunteer moderators. Overall, we show that introducing algorithm-based content moderation into platform governance is beneficial for sustaining digital communities.

History: Eric Zheng, Senior Editor; J.J. Hsieh, Associate Editor.

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/isre.2021.0036>.

**Keywords:** content moderation • human-machine collaboration • bot • volunteer moderators • platform governance

## 1. Introduction

As online platforms and communities grow, inappropriate content such as hate speech and trolling pose safety challenges and economic loss to platforms (Roberts 2014, Matias 2019a, Wu et al. 2021). Diverse and complex online environments increase the need for devising platform policies and, consequently, content moderation when content policies are violated. Typically, a small subset of community members has served an important role as volunteer (human) content moderators. Platforms such as Reddit, Wikipedia, and Stack Overflow (Zheng et al. 2019) have largely relied on volunteers to moderate their content and platform environment (Jhaver et al. 2019b, Matias 2019a). With the increasing reach and influence of digital platforms, there is a growing emphasis on more platform accountability. Volunteer moderators play an important role in meeting the challenges of creating a safe environment for all users on digital platforms.

Based on the functionality of moderation tasks, volunteer moderators play community-policing and -nurturing roles. Volunteers and platforms broadly face several challenges in content moderation. First, content moderation is expected to scale in response to the fast-growing user base and the corresponding growth in the volume of online content. It is, however, unrealistic to expect volunteer moderation capacity to scale without effective automation support. Volunteer moderators' time is not scalable, and they frequently experience burnout in handling the large volume of online content and rule violations (Grimmelmann 2015, Gillespie 2018, Dosono and Semaan 2019, Matias 2019b). Second, besides preventing the community from being exposed to inappropriate content, volunteer moderators are also expected to nurture community discourse, which emphasizes curating discussions through user education on community norms and fostering improved dialogue with community members (Yu et al. 2018, Jiang

2020, Ruckenstein and Turunen 2020). Nurturing-oriented moderation requires more moderator effort and attention to individual needs than policing-oriented moderation. Motivating desirable nurturing moderation has been a great challenge for online platforms. Last, as volunteer moderators engage in unpaid work, their retention has been the key to the sustainability of the volunteer-based platform-governance model. Platforms are motivated to seek solutions that attract and sustain volunteer engagement and contributions and minimize attrition.

The developments in algorithms and machine learning techniques open a new avenue for platforms to address the challenges to achieve scalable, desirable, and sustainable content moderation (Hammer 2016, Seering et al. 2019). For instance, Facebook has applied machine learning techniques to detect pornographic content (Roberts 2014, Gillespie 2018). On Reddit, many communities have started to embed moderation tasks and processing logic into algorithms (often called bots) to screen and remove inappropriate content (Chandrasekharan et al. 2018). Whereas platforms are moving toward the algorithm-assisted mode of content moderation, findings from researchers and reactions from the public are mixed. On the one hand, the direct interaction between volunteer moderators and users is critical for community building (Ruckenstein and Turunen 2020). Thus, bots may substitute volunteers in moderating content, leading to reduced volunteer moderation effort and engagement with community users. On the other hand, automating moderation tasks through bots may also result in an increased volunteer effort. The automation eases the workload on volunteer moderators by reducing the need to handle tedious tasks, for example, detecting rule violations and removing inappropriate content (Jhaver et al. 2019c). The reduced workload and improved work environment (e.g., less tedious work) can, in turn, enhance commitment (Smith 1994, Alfes et al. 2016). Simultaneously, volunteer moderators are better able to expand their role as community rule compliance officers, allowing them to engage in moderation that requires more subjective judgment and even offering more support to their community to nurture its members (Karusala et al. 2017).

Relevant academic research on human-bot collaboration in content moderation is still in its infancy. First, information systems researchers have long studied platform-governance models (Sambamurthy and Zmud 1999, Tiwana et al. 2010) and content moderation (Kraut and Resnick 2012, Ren and Kraut 2014). However, despite calls for research into the hybrid mode of content moderation with both human and technological effort (Ren and Kraut 2014), few studies have empirically examined the influence of machine adoption on volunteers' moderation efforts in online platforms. Second, the machine substitution and augmentation literature has

investigated various contexts, focusing predominantly on compensated workers (Autor and Dorn 2013, Acemoglu and Restrepo 2018). As an emerging context of great importance to modern society, content moderation has, however, received little attention, particularly regarding the role of automation in volunteer-based work. Third, an emerging literature on content moderation has started to provide qualitative and descriptive evidence on human and bot performance (Jhaver et al. 2019c, Ruckenstein and Turunen 2020). These studies have laid the foundation for our empirical investigation into the influences of bots on volunteer moderation.

In this study, we seek to investigate the impact of algorithm-powered bot moderators on volunteer moderators' activities. Formally, with the focus on two critical roles of volunteer moderators and the challenges in content moderation (i.e., a growing need for more scalable and nurturing moderation), we approach our research objectives by investigating two research questions: (RQ1) How do bots affect volunteer moderators' community-policing efforts? (RQ2) How do bots affect volunteer moderators' community-nurturing efforts?

To answer the research questions, we select Reddit as our research context. The Reddit platform has a plethora of communities named subreddits with each having a small number of volunteer moderators. To address the scale of the user base, Reddit has developed bots (e.g., AutoModerator) to facilitate routine content moderation. Volunteer moderators of subreddits can integrate the bots to assist in content moderation. Once integrated, when a bot detects a rule violation on Reddit, it follows the subreddit community guidelines and predefined workflow to take actions such as flagging and removing content.

We collect bot and volunteer public moderation records from 156 subreddits on Reddit from 2013 to 2014. We identify the bot-automated moderation tasks from the public moderation records. To identify volunteer moderators' activities, we employ an advanced natural language processing technique, bidirectional encoder representations from transformers (BERT). A difference-in-differences (DiD) model is then applied to estimate the impact of moderation automation. Our results suggest that bots stimulate volunteer moderators to increase their policing activities by 20.9%. Meanwhile, bots encourage volunteers to exert more community-nurturing efforts. The number of explanations created by volunteers in a month increases by 15.2% after adopting bots. And such effects are more pronounced in larger communities with greater moderation needs. We find that bots augment rather than substitute volunteer moderators by stimulating more of their policing efforts to enforce subjective rules and expand their scope of work. Moreover, the increased community-nurturing efforts are primarily driven by the increasing need to accompany policing

activities over subjective rules. Additionally, volunteer-based moderation becomes more sustainable with bots as we observe increased retention of volunteer moderators after bot implementation. We perform a series of robustness checks and validate the findings.

This research contributes to three streams of literature. We first contribute to the platform-governance literature (Sambamurthy and Zmud 1999, Tiwana et al. 2010, Ren and Kraut 2014) by investigating the hybrid mode of content moderation that integrates humans and algorithms and how the two types of moderators achieve augmentation. Our study suggests that algorithms can augment humans by stimulating more policing and nurturing moderation. Second, our work contributes to the machine substitution and augmentation literature by extending this stream of work into volunteer-based content moderation tasks (Autor and Dorn 2013, Acemoglu and Restrepo 2018, Dixon et al. 2021, Fügner et al. 2022). Our research suggests that machines complement voluntary labor by shifting their attention to tasks requiring specific responses and subjective judgment. Third, our study contributes to the growing content moderation literature (Dosono and Semaan 2019, Matias 2019b, Zheng et al. 2019) by providing novel empirical evidence on the role of bot moderators. We offer comprehensive analyses of how bots affect volunteers' activities and retention. Our research indicates that bots are an effective tool for improving the scale and quality of content moderation in online communities.

## 2. Related Literature and Hypotheses

### 2.1. Content Moderation and the Role of Volunteer Moderators

Content moderation refers to the mechanism that regulates individuals' participation in a community to increase the quantity and quality of engagement (Ren and Kraut 2014, Grimmelmann 2015). Content moderation is an important aspect of platform governance, especially for platforms that rely on user-generated content.

Central to content moderation on online platforms are volunteer moderators. Drawing on the role theory (Biddle 1986, Seering et al. 2018, Zheng et al. 2019, Tarafdar et al. 2023) and the extant literature on online communities (Kraut and Resnick 2012, West 2018, Jiang 2020, Ruckenstein and Turunen 2020, Yu et al. 2020), we conceptualize volunteer moderators' functional roles based on the tasks and associated goals into two categories: community-policing and community-nurturing. Online Appendix A summarizes related literature and categorization. Community-policing focuses on coercing compliance, such as tools used to limit inappropriate content and user behavior, and it emphasizes the speed and efficiency of content removal. In contrast, community-nurturing focuses on curating discussion by educating

users about community norms and improving dialogue with community members. Volunteer moderators are expected to perform community-policing activities to remove undesirable and harmful content because it can effectively reduce community noncompliance rates and further maintain community integrity and order (Srinivasan et al. 2019, Wu et al. 2021). In addition, they are expected to engage in community-nurturing endeavors to provide valuable support to the community (Yu et al. 2020). Such activities hold the promise of fostering more constructive interactions within the community, nurturing long-term community growth (Karusala et al. 2017, West 2018, Yu et al. 2018, Ruckenstein and Turunen 2020).

Community-policing and -nurturing are interdependent. On the one hand, more policing can elicit increasing needs for community-nurturing efforts, such as offering explanations and suggestions. On the other hand, nurturing moderation complements policing-oriented moderation in achieving more effective platform governance. For example, Jhaver et al. (2019a) find that, when executing content removal, offering sufficient explanations would significantly reduce the possibility of future rule violations.

In practice, community-nurturing moderation widely suffers from inadequate effort (West 2018, Jiang 2020, Ruckenstein and Turunen 2020) compared with community-policing moderation because of the greater cognitive complexity involved in nurturing-oriented activities. Moreover, existing literature primarily focuses on policing-oriented moderation and its influence. Little attention has been paid to volunteer moderators' dual roles in content moderation and examining potential approaches to encourage greater and more desirable community-nurturing efforts from moderators. We herein focus on whether and how bot-assisted moderation may affect volunteer moderators' efforts in policing- and nurturing-oriented moderation.

### 2.2. Bot Capability and Automation Impacts

Online platforms increasingly rely on human and algorithmic agents (e.g., bots) in content moderation (Ren and Kraut 2014). Extant literature studies the capability of human and bot moderators (Gillespie 2018, Seering et al. 2018, Jhaver et al. 2019c, Zheng et al. 2019). For example, Zheng et al. (2019) extract bot functions based on bot description and activity history from Wikipedia and then categorize bots' roles. They find that bots play both protector and advisor roles on Wikipedia. The protector role is analogous to the community-policing role, focusing on tasks such as identifying spam, vandals, and policy violations. The advisor role is analogous to the community-nurturing role, focusing on tasks such as providing suggestions for users and greeting newcomers. Notably, with a focus on bot usage in content moderation on Reddit, Jhaver et al. (2019c) find that bots

prove effective in replacing human moderators to remove inappropriate content. Nonetheless, the study also revealed deficiencies in the bots' ability to navigate situations that demand an understanding of nuanced contextual details. As a result, human moderation remains indispensable for enforcing community guidelines that necessitate subjective interpretation.

A great deal of literature also investigates machine substitution and augmentation in various contexts, such as productivity and employment (Autor and Dorn 2013, Acemoglu and Restrepo 2018, Rai et al. 2019, Dixon et al. 2021, Fügener et al. 2022). Researchers find that machines show great strength in performing routine, simple, and highly predictable tasks without requiring complex cognitive effort. Therefore, machine automation tends to substitute humans in performing such tasks. However, constrained by machine capability in cognitively complex tasks, automation may also increase the need for human efforts to perform tasks that complement automated ones. Plentiful empirical evidence indicates the substitutional and complementary impacts of machine automation on human productivity in various contexts. For example, Dixon et al. (2021) study robot adoption in the manufacturing industry and find that robots decrease the employment of middle-skilled workers (such as technicians) who follow working protocols. In contrast, robots increase the employment of high-skilled labor who can work with robots and conduct more cognitively complex tasks.

Extant literature, thus, establishes a great foundation for understanding machine capability and how automation affects labor demand and human productivity. However, prior literature mainly focuses on compensated labor; little attention has been paid to automation and its influences on content moderation that heavily relies on voluntary human effort. As granular moderation data are needed to dissect the effects of bot automation in content moderation, to this end, we collect data from a real-world setting and utilize measures of volunteers' moderation efforts to empirically investigate whether and how the adoption of algorithm-based moderation bots substitutes or complements volunteer moderators.

### 2.3. Bot Adoption and Community-Policing Moderation

Drawing on the machine substitution and augmentation literature, the impact of bot adoption on volunteers' community-policing moderation efforts is heavily contingent upon task-specific characteristics (Autor and Dorn 2013, Acemoglu and Restrepo 2018, Dixon et al. 2021). Volunteer moderators exercise their policing decisions by following respective community guidelines. These guidelines outline a series of community rules governing acceptable user-generated content, primarily focusing on content format and quality standards (Fiesler et al. 2018, Seering et al. 2019). Considering the

inherent complexity of community rules, bots could exhibit varying capabilities in automating each rule, leading to a diverse range of impacts on the workload of volunteer moderators.

In the case of simple and objective rules, bots possess the potential to efficiently automate related policing tasks on a scalable basis, thereby reducing the reliance on human effort. For instance, research demonstrates that employing bots for policing simple and objective tasks can yield comparable performance levels to those achieved by human moderators. Interviews with volunteer moderators on Reddit find that bots can effectively identify certain forms of personal attacks and other inappropriate content (Jhaver et al. 2019c). Follow-up research further reveals that, when platform moderation maintains high transparency, users perceive no distinction between removals carried out by human or bot moderators (Jhaver et al. 2019a).

In contrast, when handling complex and subjective rules that involve more cognitive evaluation and contextual judgment, bots exhibit limitations in displacing humans to exercise policing decisions (Jhaver et al. 2019c, Seering et al. 2019). Complex community rules, such as detecting satirical language and hate speech, still largely rely on human judgment; thus, bots show limited capability in automating policing-oriented moderation on subjective tasks. Moreover, as discussed earlier, when bots automate simple and objective rules, volunteer moderators have more bandwidth to enforce subjective rules. Therefore, volunteer moderators' moderation activities will not necessarily decrease after bots are introduced to the community governance process.

To summarize, bots can exhibit both substitutional and complementary effects on moderators' community-policing efforts. On the one hand, bots can substitute volunteer effort by effectively detecting and automating policing on objective tasks. On the other hand, bots can complement volunteers' efforts by enabling them to tackle policing more subjective tasks. Thus, volunteers' effort in performing community-policing moderation is determined by the net influences of substitutional and complementary effects from bot adoption. Therefore, we propose a set of hypotheses as follows.

**Hypothesis 1a.** Communities' bot adoption for content moderation leads volunteer moderators to perform more (subjective) community-policing moderation.

**Hypothesis 1b.** Communities' bot adoption for content moderation leads volunteer moderators to perform less (objective) community-policing moderation.

### 2.4. Bot Adoption and Community-Nurturing Moderation

Given the volunteer moderators' dual roles as community compliance officers and nurturers, we theorize that communities' bot adoption for content moderation has

the potential to impact volunteer moderators' efforts in community-nurturing moderation because of attention reallocation and increased need for complementary labor. Community-nurturing moderation involves volunteers' thoughtful decisions and active support for the community. When exercising nurturing-centered moderation, volunteer moderators normally need to handle various requests from community members and offer personalized and contextualized assistance. The task characteristics of community-nurturing moderation make it difficult to automate. A growing body of evidence shows that labor substitution through automation can induce demand and productivity for nonautomated tasks (Acemoglu and Restrepo 2018, Brynjolfsson et al. 2018, Dixon et al. 2021). For example, Gombolay et al. (2018) study robots and patient care in hospitals, and they find that, when hospitals adopt robots for the purpose of lifting patients from their beds, nurses are not only relieved from such physically demanding tasks, but also able to perform more caring activities such as more interactions with patients. In the moderation automation context, prior to the adoption of automated solutions for content moderation, volunteer moderators need to allocate their limited time and effort to policing and nurturing work. However, when bots automate policing activities, volunteer moderators are able to pay more attention to nurturing moderation work.

Moreover, machine augmentation literature suggests that automation may increase the demand for existing tasks that complement automated ones. It is important, therefore, to recognize that policing and nurturing moderation are interdependent. The changes in community-policing moderation after bot adoption may increase the need for complementary nurturing activities. Effective policing typically requires accompanying nurturing actions, such as justifying policing decisions, addressing individual users' special requests, and providing personalized suggestions (Yu et al. 2020). Considering the potential increase in volunteers' engagement in policing subjective rules following the adoption of bots in content moderation, as discussed in Section 2.3, this increase will prompt a greater demand for community-nurturing activities centered around policing decisions. Direct and personalized involvement by volunteer moderators is, therefore, likely required to tackle confusion and elucidate their thought process behind the decision. For example, volunteer moderators may be expected to provide subtle nudges or point to alternative solutions to community members to assist their further participation on the platform. To summarize, because of the reallocated efforts and increased need for complementary nurturing activities, we propose the following hypothesis.

**Hypothesis 2.** *Communities' bot adoption for content moderation leads volunteer moderators to perform more community-nurturing moderation.*

### 3. Empirical Setting

#### 3.1. Research Context

We collected and compiled a data set from Reddit, a large online discussion and community platform. Reddit comprises more than 130,000 active communities, termed subreddits, covering a wide range of subjects such as world news, sports, writing, and movies. Reddit users can engage with the subreddits by posting content, leaving comments, and voting on others' content on Reddit. Reddit has more than 52 million daily active users, and it has become the seventh most visited user-generated content platform in the world as of 2020.<sup>1</sup>

Reddit used to rely solely on volunteer moderators to manage all its communities. Most volunteer moderators are community users who have a deep understanding of the subreddit's community theme and rules. These volunteer moderators design community rules and monitor community activity regularly. When a rule violation happens, they are authorized to remove content and take disciplinary actions against users.

Since 2012, some subreddits have started to utilize algorithm-based moderation tools, known as bots, to assist their content moderation. Two characteristics make Reddit an ideal research context for our study: availability of moderation records and varied timings of bot adoption by different subreddits. First, because of security and intellectual property concerns, most platforms do not disclose details of the technical strategies used for their content moderation. We observe moderation records for neither human nor bot moderators for those platforms. In contrast, Reddit makes task automation and moderation records available to the public. On Reddit, researchers can identify subreddits that use bots to assist with their community governance (Chandrasekharan et al. 2018, Fiesler et al. 2018, Chandrasekharan and Gilbert 2019). Typically, these bot moderators automatically inspect newly submitted content and conduct a series of actions accordingly, such as removing offensive content (Dosono and Semaan 2019). Each bot moderator has a profile page displaying its creation date and historical moderation records. Most importantly, moderation records are presented in the form of public comments as shown in Figure 1. These comments include details such as moderation time, action, explanations, and suggestions to the content creator. Volunteer moderators perform their moderation in a similar manner. When they find a rule violation, they can perform moderation and inform the content creators through public comments, submission flairs, or private messages. Public comment is the dominant form of recording moderation on Reddit (Jhaver et al. 2019b); therefore, it is the primary data source for our study. Based on these records, we recover the moderation automation timeline and investigate its impacts on volunteer moderators' activities.

**Figure 1.** (Color online) AutoModerator's Profile Page and Moderation Records

The figure consists of two side-by-side screenshots. On the left is a screenshot of a Reddit comment from the r/LegalAdviceUK subreddit. The comment was made by u/Sharkary and has been removed by a moderator. The text of the comment is as follows:

Comment removed by moderator 9 months ago  
AutoModerator MOD Score hidden · just now  
FYI, this comment has been removed as the thread you are commenting in is an old thread. This means the information contained in the thread may be out of date, unmonitored by the community, and not likely to receive any further attention. If you are asking legal help, please consider [making a new thread](#) to receive advice.  
*I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](#) if you have any questions or concerns.*

On the right is a screenshot of the AutoModerator user profile on Reddit. The profile picture is a green shield icon. The profile information includes:

u/AutoModerator Karma 8,095,966 Cake day January 4, 2012  
+5.5k Received the Platinum Award and more in the past 30 days

Follow More Options

Trophy Case (17)

Second, the adoption timings of bot moderators in different subreddits are varied, which provides an opportunity to leverage a DiD research design. The focal bot moderator in our study is called AutoModerator.<sup>2</sup> AutoModerator was first introduced in 2012 and has been the most influential bot moderator on Reddit. More than 4,000 subreddits have adopted AutoModerator as of December 2019. The bot generates thousands of moderation records every day. To utilize AutoModerator, volunteer moderators need to follow simple syntaxes to add bot-related codes to their subreddit meta page. Common rules automated by AutoModerator are listed in Online Table C1. For example, AutoModerator can restrict users below a certain reputation level threshold on the platform. They can also prevent users from posting content from blocklisted sources. Subreddits have the flexibility to adapt and integrate AutoModerator functions to meet the subreddit guidelines and moderation needs.

### 3.2. Data Collection and Measures

Our data collection includes the following steps. First, we collected all moderation records by AutoModerator from 2013 to 2014 using the PushShift application programming interface (Baumgartner et al. 2020).<sup>3</sup> We situated our study in this observation period because AutoModerator's function was relatively stable after being released for a year. Further, more subreddits started deploying bots for moderation after 2013. Thus, the selection of subreddits in that period contained ample variations for empirical analyses. Second, we marked the time when the first AutoModerator moderation was performed in a subreddit as the moment when that specific subreddit adopted AutoModerator. With the complete moderation records collected in the first step, we observed that 156 subreddits adopted

AutoModerator from 2013 to 2014. We report the complete list of subreddits in Online Appendix B. Next, we identified volunteer moderators for each studied subreddit. Reddit grants moderators a “mod” flair to signify their role as community managers, and this flair is shown along with the moderators’ comments. Taking advantage of this platform design feature, we obtained the list of users with the mod flair and further collected their comments during the observation window. Lastly, using AutoModerator and volunteer moderators’ comments, we performed natural language processing (NLP) to extract the automated rules and volunteer moderation types.

To facilitate the analyses, we performed two data preprocessing steps. On the AutoModerator side, we extracted the automated moderation tasks and corresponding implementation time to recover the timeline of task automation. We utilized this timeline to construct the independent variable of interest *AutomatedTasks<sub>it</sub>* in our empirical analyses. Given that AutoModerator follows standardized formats for any moderation record, we employed a rule-based approach to extract automated tasks and their implementation time.

On the volunteer moderator side, our preprocessing task is to identify their moderation types. We first performed a theory-driven annotation (labeling) for a sample of volunteer moderators’ comments. Following community-policing and -nurturing moderation categorization, we first included policing as one category in our annotation. Next, we measured community-nurturing moderation with two typical nurturing activities highlighted in extant literature: explanation and suggestions (West 2018, Jhaver et al. 2019a, Jiang 2020). Following real-world practices, a moderation-related comment can have both community-policing and -nurturing components (West 2018, Jhaver et al. 2019a).

Thus, our annotation allows policing- and nurturing-oriented moderation to coexist in a volunteer moderator's comment. Additionally, we included casual talk as the fourth activity category in our annotation, which includes remaining volunteer moderators' activities besides community-policing and -nurturing ones. Table 1 illustrates example comments for each category. Compared with the preprocessing of bot comments, this task is relatively challenging because of the diversity and complexity of human language. Our methods need to go beyond the vocabulary to grasp the semantic meaning and purpose of a comment.

We utilized BERT, a deep learning approach, to classify volunteer moderators' moderation types (Devlin et al. 2018). BERT was proposed in 2018; it has quickly become the state-of-the-art NLP technique and has achieved exceptional performance in tasks such as classification and commonsense reasoning. Compared with traditional NLP methods, BERT is a bidirectional language model trained on a large-scale corpus. It has a better sense of the language context and the relationship between all words regardless of their position. Therefore, BERT can capture the meaning of comments regardless of the vocabulary choice and sentence sequence. We used BERT-based classifiers to identify volunteer moderation types.

Operationally, we used the pretrained BERT model and then fine-tuned the parameters with the labeled data set. Pretrained BERT can capture the context-free meaning of the comments, whereas the labeled data provides more details about our context. We obtained the pretrained model from Google Research, which is trained on large corpora collected from Wikipedia and book chapters.<sup>4</sup> Then, we randomly sampled 1,017 comments from our whole data set and manually labeled these comments for each category. Two graduate students in the computer science master's program at a large public research university labeled these comments after a training session with the researchers on the annotation task. They achieved 85% consistency in their labeling. For the remaining 15% of comments, they further discussed their annotation discrepancies to reach an agreement. We fed the pretrained model with the labeled data to fine-tune the parameters. We created a

classifier for each comment category, thus forming a total of four binary classifiers. Lastly, we used the fine-tuned model to predict all collected volunteer comments.

Online Table C3 shows the classification results of the four classifiers. We achieved higher than 90% on the F1 scores in three out of four classification tasks. Notably, the policing classifier reached 97.34% accuracy, 96.99% precision, 95.68% recall, and 96.33% F1 score. Compared with typical NLP classification tasks, our BERT-based classifiers achieved excellent performance.

With all the collected data, we constructed a series of variables for the corresponding data analyses. We aggregated our data at the subreddit monthly level. We used the number of comments created by volunteer moderators in each category to measure their various types of engagement. In addition to the data mentioned above, we collected the number of community users' comments from each subreddit and used it as the proxy for user engagement in that subreddit. Table 2 depicts detailed descriptions and summary statistics of variables used in this study. The correlation table is presented in Online Table C4.

## 4. Empirical Analyses and Results

### 4.1. Identification Strategy and Main Model

Considering that the automation of moderation tasks is a shock to the subreddit and different subreddits adopted bots for moderation at different times, we apply the DiD model as our identification strategy. Specifically, the treatment group comprises communities that adopted bot moderators during a given month in the observational window, whereas the control group includes communities that relied entirely on volunteer moderators in a given month. We note that the same approaches were applied in prior literature studying the impacts of ride-sharing entry (Greenwood and Wattal 2017, Babar and Burtch 2020) and technology adoption (Tan and Netessine 2020). By comparing volunteer moderators' activities in the treatment group with those in the control group, we can estimate how the introduction of bots into content moderation affects volunteer moderation efforts.

Formally, our empirical model is presented in Equation (1). Note that  $VolunteerModeration_{it}$  denotes the dependent variable. It refers to several variables that measure

**Table 1.** Examples of Different Types of Moderator Comments

Category	Subcategory	Example
Community-policing moderation	Policing	"Removed/approved."
Community-nurturing moderation	Explanation	"...because you did not include the resolution in the title and because you did not include the location in the title."
	Suggestion	"There are instructions in the FAQ as well as our other submission guidelines. Your submission would be more appropriate in r/villageporn. Feel free to resubmit with the resolution in the title. Thanks!"
Others	Casual talk	"Thank you, and it is even more beautiful in reality, especially the very special light of Iceland."

**Table 2.** Variables and Descriptive Statistics

Variable	Description	Mean	Standard deviation	Minimum	Maximum
<i>BotAfter<sub>it</sub></i>	Dummy variable. The value equals one if subreddit <i>i</i> had adopted AutoModerator in month <i>t</i> ; zero otherwise	0.616	0.487	0	1
<i>AutomatedTasks<sub>it</sub></i>	The number of moderation tasks automated by AutoModerator on subreddit <i>i</i> in month <i>t</i>	2.231	3.342	0	29
<i>Num_Policing<sub>it</sub></i>	The number of policing-type comments created by volunteer moderators on subreddit <i>i</i> in month <i>t</i>	63.882	247.128	0	4,452
<i>Policing_Scope<sub>it</sub></i>	The number of unique rules enforced by volunteer moderators on subreddit <i>i</i> in month <i>t</i>	2.426	4.322	0	40
<i>Num_Explanation<sub>it</sub></i>	The number of explanation-type comments created by volunteer moderators on subreddit <i>i</i> in month <i>t</i>	80.166	282.480	0	4,460
<i>Num_Suggestion<sub>it</sub></i>	The number of suggestion-type comments created by volunteer moderators on subreddit <i>i</i> in month <i>t</i>	84.352	256.130	0	4,634
<i>Num_Others<sub>it</sub></i>	The number of casual comments created by volunteer moderators on subreddit <i>i</i> in month <i>t</i>	215.862	446.524	0	3,711
<i>Num_Mods<sub>it</sub></i>	The number of active volunteer moderators on subreddit <i>i</i> in month <i>t</i>	3.827	4.440	0	52
<i>Num_User_Cmts<sub>it</sub></i>	The number of comments created by users on subreddit <i>i</i> in month <i>t</i>	89,351.66	358,182.9	0	5,192,588

different kinds of volunteer moderator activities, including their community-policing efforts (*Num\_Policing<sub>it</sub>*), community-nurturing efforts (*Num\_Explanation<sub>it</sub>*, *Num\_Suggestion<sub>it</sub>*), and other casual activities (*Num\_Others<sub>it</sub>*). The key independent variable is *BotAfter<sub>it</sub>*. It is a time-variant, binary variable, and the value of one means the subreddit *i* had adopted AutoModerator in month *t*. Additionally, each volunteer moderator's activity may be subject to changes in the number of active volunteer moderators and the number of submissions their subreddit receives. Thus, we add two control variables—the number of active volunteer moderators (*Num\_Mods<sub>it</sub>*) and the number of user comments (*Num\_User\_Cmts<sub>it</sub>*) in respective subreddits—to the analysis. Another source of empirical concern is the unobserved intrinsic difference among

communities and temporary shocks on the platform. We, therefore, add the subreddit fixed effects  $\theta_i$  and month fixed effects  $\delta_t$  to control the unvarying subreddit effects and common time trends;  $\varepsilon_{it}$  is the error term.

$$\begin{aligned} \text{VolunteerModeration}_{it} = & \alpha + \beta \times \text{BotAfter}_{it} \\ & + \gamma \times \text{Controls}_{it} + \delta_t + \theta_i + \varepsilon_{it}. \end{aligned} \quad (1)$$

## 4.2. Results

Table 3 reports the main estimation results. Several interesting findings are found. First, from the estimates of *BotAfter<sub>it</sub>* in the first column, we can see that volunteers' community-policing efforts significantly increased by 20.9% after their subreddit automated moderation tasks

**Table 3.** The Impact of AutoModerator on Volunteers' Moderation

Variables	<i>Community_Policing</i>		<i>Community_Nurturing</i>		<i>Others</i>
	<i>Num_Policing</i>	<i>Num_Explanation</i>	<i>Num_Suggestion</i>	<i>Num_Others</i>	
<i>BotAfter</i>	0.209** (0.090)	0.152* (0.084)	0.077 (0.071)	-0.044 (0.062)	
<i>Num_User_Cmts</i>	0.108*** (0.037)	0.116*** (0.037)	0.108*** (0.036)	0.177*** (0.038)	
<i>Num_Mods</i>	1.447*** (0.090)	1.596*** (0.083)	1.729*** (0.073)	1.916*** (0.075)	
Constant	-1.051*** (0.309)	-0.970*** (0.319)	-0.783** (0.306)	-0.628* (0.329)	
Subreddit fixed effects	Yes	Yes	Yes	Yes	
Month fixed effects	Yes	Yes	Yes	Yes	
Number of observations	3,744	3,744	3,744	3,744	
<i>R</i> <sup>2</sup>	0.805	0.835	0.858	0.904	

Notes. Cluster-robust standard errors in parentheses, clustered at the subreddit level. For all dependent variables and control variables, we use natural log transformation (i.e., ln(x + 1)) to accommodate skewed distribution and zeros.

\*\*\**p* < 0.01; \*\**p* < 0.05; \**p* < 0.1.

with AutoModerator. This result indicates that, although bots automate some policing tasks, bot automation augments rather than substitutes volunteer efforts in achieving more policing moderation. Overall, we see volunteer moderators exhibit more policing efforts, which we break down into objective and subjective rule-related policing in a later analysis in Section 5.1.1.

Second, with regard to volunteers' community-nurturing efforts, from the second column, we observe that the number of explanations volunteer moderators made in a month increased by 15.2% after adopting bots to assist community moderation. Although we do not observe significant changes in the number of suggestions created by volunteer moderators following the bot adoption under the conventional significance level, the estimated sign aligns with the changes in volunteers' explanation activities. Note that using a seemingly unrelated regressions (SUR) approach and several other estimation approaches (in the robustness check sections with tables reported in Online Appendix G), we do observe significant effects on both explanations and suggestions. Overall, we observe evidence that supports Hypothesis 2.

Moreover, estimates associated with control variables align with our expectations. For example, volunteer moderator activities increase when there are more user submissions and active volunteers in the moderation team. We also tested the incremental impact of moderation automation and performed analyses using the Poisson pseudo maximum likelihood estimator as presented in Online Appendices D and E (Burtch et al. 2018, Kummer et al. 2020). Overall, the results remain consistent.

Additionally, we performed heterogeneity analysis to examine how demand for moderation interacts with bot adoption (Demerouti et al. 2001, Bakker and Demerouti

2017, Tarafdar and Saunders 2022). Specifically, we consider the moderating role of community size (Butler et al. 2014, Ren and Kraut 2014). Compared with subreddits with fewer members, large communities usually have greater moderation demand to accommodate the scale of user-generated content and user requests. Volunteer moderators face a greater need for their time to moderate community content; thus, they are likelier to experience a shortage of moderation efforts in large communities. When algorithm-assisted moderation tools become available, adopting these tools is more likely to augment volunteers to accomplish more moderation, including policing- and nurturing-oriented ones.

To test the moderating role of community size, we proxy the community size using the number of commenters in a subreddit as of the month when AutoModerator was adopted. Next, we split all studied subreddits into two groups based on the median value of the community size and construct a binary variable  $PopSubreddit_{it}$  to denote subreddits with relatively more members. We add an interaction term between  $PopSubreddit_{it}$  and  $BotAfter_{it}$  into Equation (1). Table 4 presents the estimation results. Overall, our results suggest that the impact of bot adoption on volunteer moderators is dependent on the community size. As expected, we note that automating moderation tasks does not change moderators' policing and nurturing efforts in small communities but positively affects voluntary moderation in larger communities. Compared with small communities, volunteer moderators from larger communities achieve 53.5% more policing activities after adopting bots. Meanwhile, these volunteers contribute 53.6% and 44.1% more explanation- and suggestion-related comments to the community with the assistance of bots. We also test another type of demand for moderation—the volunteers'

**Table 4.** The Moderating Effect of Community Size

Variables	Community_Policing		Community_Nurturing		Others
	Num_Policing	Num_Explanation	Num_Suggestion	Num_Others	
BotAfter	−0.061 (0.093)	−0.118 (0.095)	−0.146 (0.090)		−0.148* (0.088)
BotAfter × PopSubreddit	0.535*** (0.167)	0.536*** (0.152)	0.441*** (0.130)		0.206 (0.129)
Num_User_Cmts	0.112*** (0.034)	0.121*** (0.035)	0.111*** (0.034)		0.179*** (0.037)
Num_Mods	1.432*** (0.091)	1.582*** (0.085)	1.718*** (0.074)		1.910*** (0.075)
Constant	−1.063*** (0.290)	−0.982*** (0.300)	−0.793*** (0.284)		−0.633* (0.321)
Subreddit fixed effects	Yes	Yes	Yes		Yes
Month fixed effects	Yes	Yes	Yes		Yes
Number of observations	3,744	3,744	3,744		3,744
R <sup>2</sup>	0.808	0.838	0.860		0.904

Notes. Cluster-robust standard errors in parentheses, clustered at the subreddit level. For all dependent variables and control variables, we use natural log transformation (i.e., ln(x + 1)) to accommodate skewed distribution and zeros.

\*\*\*p < 0.01; \*\*p < 0.05; \*p < 0.1.

scope of work—and present the results in Online Appendix F. Overall, we find the volunteers' scope of work positively moderates the impacts of bot adoption on volunteer policing and nurturing activities.

## 5. Additional Analyses

### 5.1. Mechanism Exploration

**5.1.1. Effort Allocation Between Subjective and Objective Rule Enforcement.** To explore the underlying mechanisms of the changes in volunteer moderators' policing-oriented effort after adopting bots, we first focus on the subjectivity of volunteers' enforced rules and investigate how volunteer moderators allocate their attention to enforcing objective and subjective rules. The procedure is as follows. First, two annotators manually labeled the subjectivity of community rules in each subreddit independently. Ninety percent of the labels from these two annotators were consistent. They discussed inconsistent annotations to reach a consensus. Online Table C2 presents examples of subjective and objective rules. Second, following the NLP procedure, we preprocessed the labeled rules and volunteers' moderation comments by removing the stop words and obtaining the stemmed words. Then, we calculated the cosine similarity between each labeled community rule and each volunteer's moderation comments. A higher similarity score suggests that a community rule is highly associated with the task enforced in that comment. Based on similarity scores, we identified the associated rule in each moderation comment and then labeled the enforced rule's subjectivity.

With the labeled subjectivity of moderation comments, we next aggregated volunteers' policing as objective or subjective rules-related (i.e.,  $Num\_Policing\_Obj_{it}$  and  $Num\_Policing\_Sub_{it}$ ) at the monthly level. We

additionally calculated the number of unique rules volunteers enforced each month (i.e.,  $Policing\_Scope_{it}$ ) and used that to capture volunteer moderation scope. We used all newly constructed measures as the dependent variables in Equation (1). Given the innate substitutive relationship between objective and subjective policing, we used a SUR model to account for the correlations between the error terms of the two dependent variables. We ran two sets of SUR models for bot adoption and the number of automated tasks in the respective bot in each subreddit. As attested in Table 5, there is a clear increase in subjective policing-oriented moderation and some evidence for the decrease in objective policing-oriented moderation. Whereas volunteers' moderation on objective tasks did not change significantly after adopting bots under the conventional significance level, as the number of moderation tasks automated by AutoModerator increased, objective rule-oriented policing significantly decreased. Possibly, bots could not fully automate policing over some objective rules that are not yet encoded in the algorithm because of their innate complexity; therefore, objective policing tasks may still require volunteer moderators' efforts to enforce compliance on novel rule infractions. Moreover, as the AutoModerator is adopted to automate largely objective moderation tasks, volunteers' moderation scope expands, likely to more subjective policing. Overall, we observe the significant increase in volunteers' policing efforts after adopting bots comes from enforcing subjective rules. Volunteer moderators, therefore, complement bots by accomplishing more nonautomated tasks.

To gain a deeper insight into the changes in volunteer policing, we further supplement the findings in Table 5 by performing text analysis on the actual rules enforced by the volunteer moderators, reported in Table 6.

**Table 5.** Subjective and Objective Rule Enforcement

Variables	Moderation focus (SUR)		Moderation focus (SUR)		Moderation scope
	$Num\_Policing\_Obj$	$Num\_Policing\_Sub$	$Num\_Policing\_Obj$	$Num\_Policing\_Sub$	$Policing\_Scope$
<i>BotAfter</i>	-0.017 (0.043)	<b>0.093**</b> (0.044)	—	—	<b>0.072**</b> (0.035)
<i>AutomatedTasks</i>	—	—	<b>-0.017***</b> (0.006)	<b>0.051***</b> (0.006)	—
<i>Num_User_Cmts</i>	0.059*** (0.014)	0.063*** (0.015)	0.058*** (0.014)	0.069*** (0.015)	0.067*** (0.023)
<i>Num_Mods</i>	0.592*** (0.032)	0.815*** (0.033)	0.605*** (0.032)	0.778*** (0.033)	0.523*** (0.041)
Constant	-0.646*** (0.233)	-0.382 (0.240)	-0.661*** (0.233)	-0.343 (0.238)	-0.534** (0.205)
Subreddit fixed effects	Yes	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes	Yes
Number of observations	3,744	3,744	3,744	3,744	3,744
$R^2$	0.685	0.759	0.685	0.763	0.778

Notes. SUR for the moderation focus analysis; standard errors in parentheses. For all dependent variables and control variables, we use natural log transformation (i.e.,  $\ln(x + 1)$ ) to accommodate skewed distribution and zeros.

\*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$ .

**Table 6.** Rule-Level Exploration

Group	Example rules in the group	Volunteers' policing activities before using bots.	Volunteers' policing activities after using bots.	<i>t</i> -statistic	<i>p</i> -value
		Mean (Std.)	Mean (Std.)		
Group 1: Reduced volunteer moderation (99 rules)	<ul style="list-style-type: none"> <li>Content that contains prohibited links or keywords</li> <li>Content with incorrect title format</li> </ul>	112.272 (49.454)	21.818 (9.009)	1.800	0.074
Group 2: Increased volunteer moderation (50 rules)	<ul style="list-style-type: none"> <li>Content with personal information in texts or images</li> <li>Content with nonconcise title and low quality</li> </ul>	56.720 (22.239)	201.120 (75.344)	-1.838	0.069
Group 3: Emerging volunteer moderation (85 rules)	<ul style="list-style-type: none"> <li>Content that may cause unexpected consequences and damage community quality and integrity, such as satire content, vote manipulation, and posts seeking medical advice</li> </ul>	0.000 (0.000)	42.127 (15.005)	-2.808	0.006

Evidence from Tables 5 and 6 collectively lends support for Hypotheses 1a and 1b.

With the identified rule in each moderation comment, we aggregated volunteers' policing on each rule before and after utilizing bots. We find three groups of rules with different patterns. Specifically, the first group in Table 6 comprises enforced rules and associated policing tasks with decreased volunteer efforts, such as removing content that contains prohibited keywords and content with incorrect title format. These rules are mostly simple and objective; thus, bots are likely to detect rule violations with a high accuracy rate. Overall, the decreased volunteer efforts on these rules show evidence of the machine substitution effect. Bots substitute volunteers in policing such tasks and reduce their associated effort.

The second group in Table 6 consists of rules that involve increased volunteer moderators' policing efforts after adopting bots. Example policing tasks in this group include removing personal information in texts/images and low-quality content (e.g., unconcise titles and content). These tasks are associated with a mix of subjective and objective rules that are difficult to code with simple and consistent patterns. The rule violations can occur in diverse content types (e.g., disclosing personal information through texts, images, or linked websites), and it is difficult to detect the violation automatically. Enforcing these rules requires more subjective judgment and a better understanding of the contextual details. The increased volunteer policing efforts on these rules imply the machine augmentation effects. Automation augments volunteer moderators by enabling them to take on more nonautomated tasks.

The third group in Table 6 shows emerging tasks volunteer moderators did not enforce until utilizing bots in their subreddit. Rules and associated policing tasks in

this group mainly focus on community content quality and integrity, such as removing content that seeks medical advice, satire content, and vote manipulations. The violation of these rules may be in various forms. Enforcing these rules usually requires a good understanding of the content and more cognitive effort. As such, it is challenging to fully automate the moderation of these rules. Interestingly, we observe this group also includes some bot-automated tasks, such as removing spam and memes. Researchers present qualitative evidence showing that automatic moderation tools such as AutoModerator can stimulate volunteer moderator efforts to manage rules that they did not tackle in the past and address potential errors caused by automated moderation (Jhaver et al. 2019c). The increased volunteer efforts in the emerging rules are also consistent with extant machine augmentation literature (Acemoglu and Restrepo 2018, Dixon et al. 2021), indicating that automation can lead to increased human labor because of the growing need for new tasks that complement automation.

### 5.1.2. Complementary Role of Community-Nurturing to Community-Policing.

As discussed in Section 2.4, two key factors might underlie the increased community-nurturing efforts by volunteers subsequent to bot implementation: (1) the attention reallocation effect from automated policing on objective rules and (2) the increased need for complementary nurturing-oriented moderation for subjective rule enforcement. To examine which mechanism drives the results in Table 3, we construct two sets of measures based on the interplay between community-nurturing and -policing activities. The first set of measures (i.e., *Num\_Explanation\_Alone*, *Num\_Suggestion\_Alone*) indicates the number of stand-alone nurturing-oriented comments that do not occur

together with policing activities; these measures capture organic and proactive volunteer efforts to build and support their communities. In contrast, the second set of measures (i.e., *Num\_Explanation\_After\_Policing*, *Num\_Suggestion\_After\_Policing*) indicate the number of nurturing-oriented comments created to accompany policing activities; these measures capture the complementary role of community-nurturing to policing. If the effort reallocation from automating objective rules by bots drives the increased community-nurturing efforts, we expect to see increased proactive, stand-alone nurturing-oriented moderation. If the increased need for the complementary nurturing effort for the increased (subjective) policing induces the results, we expect more increased reactive nurturing activities following policing actions.

We used the newly constructed measures as the dependent variables in Equation (1) and performed the analyses again. We observed several interesting results in Table 7. We did not see significant changes in volunteer moderators' proactive nurturing efforts in offering suggestions to community members (last column). However, considering the interdependence between policing and nurturing, we find significant increases in volunteer nurturing activities accompanying policing decisions. The results in the first two columns indicate the number of explanations and suggestions following policing activities increased by 15.9% and 16.2% after bot adoption, respectively. Overall, the results in Table 8 suggest that the increased community-nurturing efforts are primarily driven by the complementary role of nurturing to policing-oriented moderation.

Zooming in on the nurturing efforts surrounding policing decisions, we further differentiated these nurturing efforts based on the subjectivity of the enforced rules. Specifically, we generated more granular measures to calculate the number of nurturing-related comments following the policing actions for subjective (i.e.,

*Num\_Explanation\_After\_Sub\_Policing*, *Num\_Suggestion\_After\_Sub\_Policing*) and objective rule (i.e., *Num\_Explanation\_After\_Obj\_Policing*, *Num\_Suggestion\_After\_Obj\_Policing*) enforcement. Considering that policing over subjective rules involves a more complex decision process and human discretion compared with policing over objective rules, we expect that the increased nurturing activities from volunteer moderators primarily come from the enforcement of subjective rules. We again estimated Equation (1) with these newly created measures as dependent variables. The results are displayed in Table 8. As expected, the results show that the increased nurturing efforts are associated with policing activities related to subjective rather than objective rule enforcement.

In addition, we perform text analysis on volunteers' nurturing-related comments to gain a deeper understanding of how volunteers contribute to community building. We categorize volunteers' nurturing activities into different groups and list the main focus of each group in Table 9. Regarding explanation, we find that a vast majority of the explanations focus on (1) explaining rules and policing decisions; (2) answering users' individual requests, in particular, requests for more elaboration on policing decisions; and (3) clarifying the community governance system and moderation procedure. Overall, the explanations supplement policing decisions by enhancing the transparency of community rules and content moderation.

Regarding volunteer suggestions, we find they primarily focus on offering personalized guidance to users after their content gets removed. For example, moderators recommend alternative communities to users if their originally posted content does not fit the community theme. Moreover, moderators offer suggestions to answer users' specific requests, such as addressing account suspension and navigating the resources on the platform. Moderators also offer suggestions proactively by sending users reminders. These reminders aim to

**Table 7.** The Interplay Between Policing and Nurturing-Oriented Moderation

Variables	<i>Num_Explanation_After_Policing</i>	<i>Num_Suggestion_After_Policing</i>	<i>Num_Explanation_Alone</i>	<i>Num_Suggestion_Alone</i>
<i>BotAfter</i>	0.159*** (0.051)	0.162*** (0.053)	0.147* (0.083)	0.068 (0.071)
<i>Num_User_Cmts</i>	0.023 (0.022)	0.009 (0.022)	0.113*** (0.038)	0.106*** (0.036)
<i>Num_Mods</i>	0.576*** (0.059)	0.595*** (0.059)	1.571*** (0.084)	1.712*** (0.074)
Constant	-0.376** (0.190)	-0.263 (0.191)	-0.944*** (0.319)	-0.770** (0.304)
Subreddit fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
Number of observations	3,744	3,744	3,744	3,744
<i>R</i> <sup>2</sup>	0.698	0.700	0.833	0.856

Notes. Cluster-robust standard errors in parentheses, clustered at the subreddit level. For all dependent variables and control variables, we use natural log transformation (i.e., ln(x + 1)) to accommodate skewed distribution and zeros.

\*\*\*p < 0.01; \*\*p < 0.05; \*p < 0.1.

**Table 8.** Nurturing-Oriented Moderation Following Policing Actions

Variables	<i>Num_Explanation_After_Sub_Policing</i>	<i>Num_Explanation_After_Obj_Policing</i>	<i>Num_Suggestion_After_Sub_Policing</i>	<i>Num_Suggestion_After_Obj_Policing</i>
<i>BotAfter</i>	0.102** (0.042)	0.001 (0.020)	0.109** (0.042)	-0.001 (0.022)
<i>Num_User_Cmts</i>	0.017 (0.019)	0.012 (0.008)	0.010 (0.017)	0.015* (0.009)
<i>Num_Mods</i>	0.337*** (0.054)	0.095*** (0.019)	0.320*** (0.049)	0.104*** (0.023)
Constant	-0.228 (0.162)	-0.114 (0.071)	-0.160 (0.144)	-0.154* (0.080)
Subreddit fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
Number of observations	3,744	3,744	3,744	3,744
<i>R</i> <sup>2</sup>	0.665	0.445	0.689	0.383

Notes. Cluster-robust standard errors in parentheses, clustered at the subreddit level. For all dependent variables and control variables, we use natural log transformation (i.e.,  $\ln(x + 1)$ ) to accommodate skewed distribution and zeros.

\*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$ .

stress community rules and facilitate users' rule compliance. Together with our results in Table 7, we see that nurturing-oriented moderation complements policing and assists moderators in building a more transparent and supportive community environment.

## 5.2. Moderator Retention

Further, we investigate a downstream outcome of volunteer moderators' effort change. Specifically, centered around the sustainability concern of online platforms primarily relying on volunteers for content moderation (Dosono and Semaan 2019, Matias 2019b, Zheng et al. 2019), we further examine how bot adoption affects volunteer moderator retention in a community. A broad range of management and volunteerism literature suggests several work-related factors that influence voluntary labor retention (Griffeth et al. 2000, Cuskelly et al. 2006, Hidalgo and Moreno 2009). Based on the changes bots bring to volunteer moderation, we expect that bots can positively affect volunteer moderator retention in the community for several reasons. First, when bots replace volunteers in processing simple objective

policing tasks, it reduces the volunteers' overall workload and enables them to perform a broader range of moderating tasks. Prior literature suggests that nonrepetitive tasks (Hidalgo and Moreno 2009) and nonexcessive workload (Griffeth et al. 2000) contribute to employees' and volunteers' intention to remain in the organization. Second, as discussed in Section 2.4, the relief from routine tasks may enable volunteer moderators to engage in more community-nurturing activities, such as responding to community members' individual requests and achieving more fulfilling work. Tasks that are gratifying and benefit others can positively influence volunteer retention (Hidalgo and Moreno 2009). Given the increased positive interaction with users and engagement in community-building activities, volunteers' moderation would have a higher intention to stay in their moderator role.

We empirically examine how bots affect volunteer moderator retention in a subreddit. We reconstructed our data as the moderator-month panel and labeled moderators who stopped performing any type of moderation in three consecutive months as exiting the moderator team.

**Table 9.** Example Subjective and Objective Rules

Nurturing type	Summary	Example moderator comment
Explanations	Explain rules and policing decisions Answer user requests and elaborate the rules	Hate speech, bigotry, and personal attacks are not allowed. Even if you give credit to the original post, it is still a repost; against the rules.
	Address confusion about governance and moderation procedure	We did not delete anything. The user deleted it.
Suggestions	Recommend alternative communities	A good home for this question is r/AskScienceDiscussion. Try posting there!
	Answer user requests	You need to contact the administrator about your account; message r/reddit.com ASAP.
	Offer detailed guidance to users after policing	Just let me know when you edited the textbox and be aware of using the "edit" and not "delete" button.
	Send reminder	Please remember to keep responses on-topic and free of jokes or speculation.

The average length of moderators' retention in the studied communities is more than nine months (the standard deviation is about seven months). We use three different models to conduct the retention analysis. The first column in Table 10 shows the results using ordinary least squares (OLS) estimation. Here, we can see that volunteers remain in the moderator team for approximately six more months after automating tasks through bots. As shown in the second column of Table 10, we also observe consistent results using the Poisson regression. We further verify our results by utilizing survival analysis. The estimation of the survival analysis implies the risk of leaving the moderator team, and we can see that automating moderation reduces the hazard rate of exit. In other words, volunteer moderators are more likely to stay in the moderation team after the community adopts bot moderators. Overall, our results suggest that automating moderation tasks can positively influence volunteer moderators' retention in the subreddit and make volunteer-based content moderation more sustainable.

## 6. Robustness Checks

We conduct several robustness checks that combine quantitative and qualitative approaches. Table 11 summarizes all tests and findings. We present the analysis description and the results in Online Appendix G. Specifically, we conduct a relative time model to examine the parallel trend assumption for the main model specification. Because communities gradually adopted the AutoModerator during the observational period, we apply alternative model estimations (two-stage DiD and doubly robust DiD) to validate results. Further, we apply several alternative methods, such as lookahead matching and instrumental variable approaches, to address the general endogeneity concern in our analyses. Last, to eliminate the concern that the changes in volunteer moderator behavior after adopting AutoModerator are caused by external factors, such as community requests, we interviewed several experienced moderators on Reddit. We confirmed that the moderators' contributions remain voluntary without any mandatory schedule or plans. Overall, the results from various robustness checks remain consistent with our main analyses.

**Table 10.** The Impact of AutoModerator on Volunteer Moderator Retention

Variables	Retention month (OLS)	Retention month (Poisson)	Survival model (Cox)
BotAfter	6.459*** (0.804)	1.066*** (0.173)	-1.629*** (0.158)
Constant	3.055*** (0.703)	1.255*** (0.164)	—
Subreddit fixed effects	Yes	Yes	Yes
Number of observations	1,752	1,752	1,174

Note. Robust standard errors in parentheses; for OLS and Poisson, standard errors are clustered at the subreddit level.

\*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$ .

## 7. Discussions

### 7.1. Key Findings

In this study, we focus on the increasing adoption of algorithm-based moderators (bots) into content moderation in online platforms. This research identifies the focus of moderation activities to be centered primarily on policing and nurturing. In both the activities, there is a certain amount of routineness or context complexity embedded into the tasks that moderators handle. Our findings show how bot adoption shifts and expands volunteer moderators' efforts from routine policing to a broader nurturing-oriented moderation activities. Combining econometric analyses of archival moderation records from 156 communities on Reddit and qualitative interviews with the community moderators, we find that incorporating bots augments volunteer moderators and contributes to satisfying the increased needs for community-nurturing activities following the policing actions. Furthermore, our results suggest that bot moderators lead to more sustainable volunteer-based governance as it increases volunteer moderators' retention in a community.

### 7.2. Theoretical Implications

Our work contributes to three streams of research. First, we contribute to platform governance literature (Sambamurthy and Zmud 1999, Tiwana et al. 2010). Whereas prior literature discusses the role of technology in platform governance (Perscheid et al. 2020) and community management (Kraut and Resnick 2012, Ren and Kraut 2014), our work empirically investigates an emerging and yet understudied type of platform-governance model on digital platforms: human-machine collaborative content moderation. Based on the functionality behind the moderation (West 2018, Jiang 2020, Ruckenstein and Turunen 2020), our study differentiates moderation and moderators' roles as policing- and nurturing-oriented. Our results suggest that such differentiation is helpful for gaining more insights into the dynamics of moderation activities. In contrast to the prevailing literature's view of policing and nurturing as independent activities, however, our study points to the need for viewing these as interrelated activities. As such, bots' automation of policing over simple and objective tasks leads to greater needs

**Table 11.** Summary of Robustness Tests

Concerns	Test	Finding	Additional details
Preadoption parallel trend assumption	Relative time model	Results remain consistent	Online Table G1 Online Figure G1
The correlation between error terms in analyses related to four types of volunteer moderator activity	Seemingly unrelated regression model	Results remain consistent	Online Table G2
Whether the changes in volunteer moderators' activity are driven by time-unvarying and unobserved community factors	Lookahead matching	Results remain consistent	Online Tables G3 and G4
General endogeneity concerns (e.g., unobserved motivation to adopt AutoModerator, unobserved confounding factors)	Instrument variables Doubly robust DiD Two-stage DiD	Results remain consistent	Online Tables G5–G8
The measurement errors in the dependent variables derived from the BERT-based classification			
Whether the estimates are robust			
Whether changes in volunteer moderators' behavior after the adoption of Automoderator are strategically managed by the community	Interviews with volunteer moderators	Volunteers' moderation were always voluntary	Available upon request

for volunteers' corresponding nurturing efforts. Viewing them as independent activities may neglect nuanced changes in bots' and volunteers' moderation focus and contributes to an incomplete understanding of augmentation benefits.

Second, our work also sheds light on machine substitution and augmentation research (Autor and Dorn 2013, Acemoglu and Restrepo 2018, Rai et al. 2019, Dixon et al. 2021, Fügner et al. 2022, Tarafdar et al. 2023). This line of research suggests that task characteristics are one of the key factors for such relationships. Our study extends the human-machine collaboration literature from contexts centering on compensated labor to online content moderation that generally rely on voluntary labor. In our context, bots augment rather than substitute volunteer moderators in both policing and nurturing activities. However, by leveraging the interplay between moderators' functional roles (Seering et al. 2018, Tarafdar et al. 2023) and the subjectivity of enforced rules, we show that both substitutional and augmentative relationships can manifest. Our proposed dimensions of moderator role (policing versus nurturing) and task subjectivity (subject versus objective) is a valuable analytic frame for future content moderation studies.

Third, our work also contributes to the growing content moderation literature. There has been increasing attention surrounding human-bot collaborative moderation and the sustainability of volunteer-based governance in recent years (Dosono and Semaan 2019, Matias 2019b, Zheng et al. 2019). Prior work, however, primarily focuses on the qualitative perspectives of human and bot roles. We extend content moderation research by providing quantitative empirical evidence confirming that bots and human moderators tackle different content moderation tasks. Most importantly, we find bots

can serve as a technological solution to motivate more volunteer-driven and desirable community-nurturing efforts, leading to more sustainable volunteerism in platform governance.

### 7.3. Practical Implications

Our work shows several aspects of practical implications. First, for online platform managers, our research suggests that algorithm-based content moderation tools are critical and effective in coping with the scale and sustainability challenges, particularly for those heavily relying on volunteer moderation. On the content management side, equipping volunteer moderators with more automatic moderation tools will stimulate volunteers to perform more desirable community-nurturing activities in addition to greater rule enforcement. On the volunteer side, the assistance of automated moderation tools enables them to cope with a larger scope of work and extend their retention in the moderator role.

Moreover, our results suggest that online community managers should better design their community rules and moderation tasks to utilize the limited attention of volunteer moderators. For example, by creating and separating moderation tasks based on their subjectivity, community managers can more effectively identify tasks suitable for automation and better direct volunteer moderators' attention to tasks that require human discretion.

Further, given the increasing need for volunteer moderators' community-nurturing efforts surrounded by policing decisions, in addition to policing-oriented bots, platform designers should pay attention to users' follow-up requests and offer more automatic tools to assist volunteer moderators' community-nurturing activities. For example, platform designers can provide bots that can point users to potentially relevant community rules or

alternative community resources following the enforcement of subjective rules. Such nurturing-oriented bots would help further reduce the demand for volunteer moderators' nurturing efforts and create a more supportive community environment.

#### 7.4. Limitations and Future Research Opportunities

We acknowledge several limitations of our research. First, moderation records on Reddit are stored in three ways: public moderation comments, submission flair, and private conversations with content creators (Jhaver et al. 2019a). Given the data availability, our data only comes from public moderation records. We, therefore, may miss some moderation recorded in submission flair and private messages. However, we do not consider this to be a severe issue in our study for several reasons. First, more than 80% of moderation records are present in the form of public comments (Jhaver et al. 2019a). Second, public moderation records represent an essential dimension of platform governance: transparency. As calls for transparent content moderation continue to be raised (Jhaver et al. 2019a), public moderation will become the dominant regulation approach on Reddit. Therefore, studying the impacts of machines on transparent moderation is also meaningful for platforms. Meanwhile, we also open this avenue for future researchers to extend the work if the private moderation records are accessible. It would also be interesting to investigate the impacts of bots on moderators' choice of communication channels for content moderation.

Second, our study only investigates the human-machine collaboration from the angle of changes in volunteers' moderation, which represents the direct interaction with community members. We did not consider the new effort and time that volunteer moderators spent creating and maintaining this collaboration, such as designing bots, fixing errors, and updating bots (Jhaver et al. 2019c). However, these costs occur in the backend and are unfortunately impossible for us to observe in this context. Nevertheless, such backstage effort would be an intriguing topic for future research. It will generate valuable insights into moderation teams' work design and management.

Third, whereas Reddit has become a giant online platform with a large user base, other platforms may have different platform regulation structures, and they may apply different approaches to integrate algorithms into daily content moderation (Zheng et al. 2019). Therefore, it is necessary to go beyond our research context and assess the external validity of the findings using data from other platforms.

Several important directions for future research have recently emerged with the adoption of more powerful techniques such as large language models (LLMs). These

machine learning models show improved capabilities in further augmenting responses to context-specific communications. Empowering bot moderators through more advanced techniques such as LLMs may further accentuate the interaction phenomenon between the two dimensions of moderator role (policing versus nurturing) and task subjectivity (subject versus objective). LLMs significantly expand machines' capabilities and yet have some limitations in handling certain tasks (OpenAI 2023); the applications of LLMs for content moderation will be an important future research direction for this research stream.

Further, the emergence of LLMs and other generative AI models enables rapid scale-up in the production of AI-generated content, which adds more complexity and challenges to platform content governance when LLMs may potentially play both content producer and content moderator roles in online platforms. As this phenomenon evolves, platform governance will need to define and establish mechanisms for fair and ethical conduct. Theoretical framework for such environments would need to distinguish between identified/unidentified machine-generated content and its policing. In other words, the dimensionality and complexity of the task space will increase dramatically with the prevalence of LLM usage in online platforms.

Last but not least, in addition to machine substitution and augmentation, AI alignment perspective will need to be introduced to understand the interplay between bots and volunteer moderators in policing- and nurturing-related activities (Gabriel 2020). Encoding community-policing and -nurturing goals into the design of bots may help achieve greater alignment between bots and volunteers, leading to reduced volunteer efforts in content moderation. In contrast, failures of AI alignment may contribute to increased demand for the volunteer moderators' times and harm the platform's ability to provide a nurturing environment. In essence, AI alignment may offer us a new theoretical frame to understand how to coordinate the interests of different parties in the platforms, such as content producers, consumers, and moderators, and guide better governance design. We hope our study paves the way for future researchers in these emerging contexts, and we anticipate development of deeper insights into effective human-machine collaborative governance across a diverse set of platforms.

#### Acknowledgments

The authors thank the senior editor, Eric Zheng; the associate editor, J.J. Po-An Hsieh; and anonymous reviewers for their insightful comments and constructive feedback, which has significantly improved the quality of the work. The authors also appreciate the valuable feedback received from seminars at Erasmus University Rotterdam and the University of Wisconsin-Madison.

## Endnotes

- <sup>1</sup> See <https://www.redditinc.com/blog/reddits-2020-year-in-review/>. Accessed August 10, 2023.
- <sup>2</sup> See <https://www.reddit.com/wiki/automoderator/>. Accessed August 10, 2023.
- <sup>3</sup> See <https://pushshift.io/>. Accessed August 10, 2023.
- <sup>4</sup> See <https://github.com/google-research/bert>. Accessed August 10, 2023.

## References

- Acemoglu D, Restrepo P (2018) Artificial intelligence, automation, and work. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 197–236.
- Alfes K, Shantz A, Bailey C (2016) Enhancing volunteer engagement to achieve desirable outcomes: What can non-profit employers do? *VOLUNTAS Internat. J. Voluntary Nonprofit Organ.* 27(2):595–617.
- Autor DH, Dorn D (2013) The growth of low-skill service jobs and the polarization of the US labor market. *Amer. Econom. Rev.* 103(5):1553–1597.
- Babar Y, Burtch G (2020) Examining the heterogeneous impact of ride-hailing services on public transit use. *Inform. Systems Res.* 31(3):820–834.
- Bakker AB, Demerouti E (2017) Job demands–resources theory: Taking stock and looking forward. *J. Occupational Health Psych.* 22(3):273–285.
- Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J (2020) The Pushshift Reddit data set. *Proc. 14th Internat. AAAI Conf. Web Social Media*, vol. 14 (AAAI, Washington, DC), 830–839.
- Biddle BJ (1986) Recent developments in role theory. *Annual Rev. Sociol.* 12(1):67–92.
- Brynjolfsson E, Mitchell T, Rock D (2018) What can machines learn, and what does it mean for occupations and the economy? *AEA Papers Proc.* 108:43–47.
- Burtch G, Carnahan S, Greenwood BN (2018) Can you gig it? An empirical examination of the gig economy and entrepreneurial activity. *Management Sci.* 64(12):5497–5520.
- Butler BS, Bateman PJ, Gray PH, Diamant EI (2014) An attraction-selection–attrition theory of online community size and resilience. *Management Inform. Systems Quart.* 38(3):699–729.
- Chandrasekharan E, Gilbert E (2019) Hybrid approaches to detect comments violating macro norms on Reddit. Preprint, submitted April 7, <https://arxiv.org/abs/1904.03596>.
- Chandrasekharan E, Mattia S, Jhaver S, Charvat H, Bruckman A, Lampe C, Eisenstein J, Gilbert E (2018) The internet’s hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proc. ACM Human-Comput. Interaction*, vol. 2 (CSCW) (ACM, New York), 32.
- Cuskelly G, Taylor T, Hoye R, Darcy S (2006) Volunteer management practices and volunteer retention: A human resource management approach. *Sport Management Rev.* 9(2):141–163.
- Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB (2001) The job demands-resources model of burnout. *J. Appl. Psych.* 86(3): 499–512.
- Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint, submitted October 11, <https://arxiv.org/abs/1810.04805>.
- Dixon J, Hong B, Wu L (2021) The robot revolution: Managerial and employment consequences for firms. *Management Sci.* 67(9): 5586–5605.
- Dosono B, Semaan B (2019) Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. *Proc. Conf. Human Factors Comput. Systems* (ACM, New York).
- Fiesler C, McCann J, Frye K, Brubaker JR (2018) Reddit rules! Characterizing an ecosystem of governance. *Proc. 12th Internat. AAAI Conf. Web Social Media*, vol. 12 (AAAI, Washington, DC), 72–81.
- Fügner A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Inform. Systems Res.* 33(2):678–696.
- Gabriel I (2020) Artificial intelligence, values, and alignment. *Minds Machines* 30(3):411–437.
- Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, New Haven, CT).
- Gombolay M, Yang XJ, Hayes B, Seo N, Liu Z, Wadhwania S, Yu T, Shah N, Golen T, Shah J (2018) Robotic assistance in the coordination of patient care. *Internat. J. Robotics Res.* 37(10):1300–1316.
- Greenwood BN, Wattal S (2017) Show me the way to go home: An empirical investigation of ride-sharing and alcohol-related motor vehicle fatalities. *Management Inform. Systems Quart.* 41(1): 163–187.
- Griffeth RW, Hom PW, Gaertner S (2000) A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *J. Management* 26(3):463–488.
- Grimmelmann J (2015) The virtues of moderation. *Yale J. Law Tech.* 17:42.
- Hammer HL (2016) Automatic detection of hateful comments in online discussion. *Proc. Internat. Conf. Indust. Networks Intelligent Systems* (Springer International Publishing, New York), 164–173.
- Hidalgo MC, Moreno P (2009) Organizational socialization of volunteers: The effect on their intention to remain. *J. Community Psych.* 37(5):594–601.
- Jhaver S, Bruckman A, Gilbert E (2019a) Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. *Proc. ACM Human-Comput. Interactions*, vol. 3 (CSCW) (ACM, New York), 150.
- Jhaver S, Appling DS, Gilbert E, Bruckman A (2019b) Did you suspect the post would be removed? Understanding user reactions to content removals on Reddit. *Proc. ACM Human-Comput. Interactions*, vol. 3 (CSCW) (ACM, New York), 192.
- Jhaver S, Birman I, Gilbert E, Bruckman A (2019c) Human-machine collaboration for content regulation: The case of Reddit Auto-moderator. *ACM Trans. Comput. Human Interactions* 26(5): 1–35.
- Jiang J (2020) Toward a multi-stakeholder perspective for improving online content moderation. Unpublished doctoral dissertation, University of Colorado, Boulder.
- Karusala N, Vishwanath A, Kumar A, Mangal A, Kumar N (2017) Care as a resource in underserved learning environments. *Proc. ACM Human Comput. Interactions*, vol. 1 (CSCW) (ACM, New York), 104.
- Kraut RE, Resnick P (2012) *Building Successful Online Communities: Evidence-Based Social Design* (MIT Press, Cambridge, MA).
- Kummer M, Slivko O, Zhang X (2020) Unemployment and digital public goods contribution. *Inform. Systems Res.* 31(3):801–819.
- Matias JN (2019a) The civic labor of volunteer moderators online. *Soc. Media Soc.* 5(2):1–12.
- Matias JN (2019b) Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proc. Natl. Acad. Sci. USA* 116(20):9785–9789.
- Open AI (2023) GPT-4 technical report. OpenAI, San Francisco.
- Perscheid G, Ostern NK, Moermann J (2020) Determining platform governance: Framework for classifying governance types. *Proc. Internat. Conf. Inform. Systems* (AIS, Atlanta).
- Rai A, Constantinides P, Sarker S (2019) Next generation digital platforms: Toward human-AI hybrids. *Management Inform. Systems Quart.* 43(1):iii–ix.
- Ren Y, Kraut RE (2014) Agent-based modeling to inform online community design: Impact of topical breadth, message volume,

- and discussion moderation on member commitment and contribution. *Human Comput. Interactions* 29(4):351–389.
- Roberts ST (2014) Behind the screen: The hidden digital labor of commercial content moderation. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.
- Ruckenstein M, Turunen LLM (2020) Re-humanizing the platform: Content moderators and the logic of care. *New Media Soc.* 22(6):1026–1042.
- Sambamurthy V, Zmud RW (1999) Arrangements for information technology governance: A theory of multiple contingencies. *Management Inform. Systems Quart.* 23(2):261–290.
- Seering J, Flores JP, Savage S, Hammer J (2018) The social roles of bots: Evaluating the impact of bots on discussions in online communities. *Proc. ACM Human Comput. Interactions*, vol. 2 (CSCW) (ACM, New York), 157.
- Seering J, Wang T, Yoon J, Kaufman G (2019) Moderator engagement and community development in the age of algorithms. *New Media Soc.* 21(7):1417–1443.
- Smith DH (1994) Determinants of voluntary association participation and volunteering: A literature review. *Nonprofit Voluntary Sector Quart.* 23(3):243–263.
- Srinivasan KB, Danescu-Niculescu-Mizil C, Lee L, Tan C (2019) Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community. *Proc. ACM Human Comput. Interactions*, vol. 3 (CSCW) (ACM, New York), 163.
- Tan TF, Netessine S (2020) At your service on the table: Impact of tabletop technology on restaurant performance. *Management Sci.* 66(10):4496–4515.
- Tarafdar M, Saunders C (2022) Remote, mobile, and blue-collar: ICT-enabled job crafting to elevate occupational well-being. *J. Assoc. Inform. Systems* 23(3):707–749.
- Tarafdar M, Page X, Marabelli M (2023) Algorithms as co-workers: Human-algorithm role interactions in algorithmic work. *Inform. Systems J.* 33(2):232–267.
- Tiwana A, Konsynski B, Bush AA (2010) Research commentary—Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Inform. Systems Res.* 21(4):675–687.
- West S (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media Soc.* 20(11):4366–4383.
- Wu J, Zheng Z, Zhao JL (2021) FairPlay: Detecting and deterring online customer misbehavior. *Inform. Systems Res.* 32(4):1323–1346.
- Yu B, Spiel K, Watts L (2018) Supporting care as a layer of concern: Nurturing attitudes in online community moderation. *Sociotechnical Systems Care CSCW18 Workshop (Hudson River, NY)*.
- Yu B, Seering J, Spiel K, Watts L (2020) Taking care of a “fruit tree”: Nurturing as a layer of concern in online community moderation. *Proc. Conf. Human Factors Comput. Systems* (ACM, New York).
- Zheng L, Albano CM, Vora NM, Mai F, Nickerson JV (2019) The roles bots play in Wikipedia. *Proc. ACM Human Comput. Interactions*, vol. 3 (CSCW) (ACM, New York), 215.