The Effects of Machine-powered Content Moderation: An Empirical Study on Reddit

Qinglai He
University of Wisconsin-Madison
qinglai.he@wisc.edu

Yili (Kevin) Hong
University of Houston
yilihong@uh.edu

T. S. Raghu
Arizona State University
Raghu.Santanam@asu.edu

Abstract

With increasing participation in social media and online communities, content moderation has become an important part of the online experience. Volunteer moderators have been the essential workforce for content moderation. Recently, platforms move toward the technical and automated mode of moderation. There is a growing concern over de-humanization and whether machines would lead volunteer moderators to reduce their contributions. We conduct an empirical study to examine the impact of machine-powered regulations on volunteer moderators' behaviors. With data collected from 156 subreddits on Reddit, we found that delegating moderation to machines augments volunteer moderators' role as community managers. Human moderators engage in more moderation-related activities, including 20.2% more corrective and 14.9% supportive activities with their community members. Importantly, the effect manifests primarily among communities with large user bases and detailed guidelines, suggesting that community needs for moderation are the key factors driving more voluntary contributions in the presence of bot moderators.

1. Introduction

Dominant online platforms have been facing safety challenges and economic loss caused by inappropriate content such as hate speech and trolling [1, 2]. Diverse and complex online environments increase the need for devising platform policies, and consequently, content moderation when content policies are violated. Platforms, such as Reddit and Wikipedia, heavily rely on human labor to maintain their content [1,3,4,5]. However, volunteer moderators usually experience burnout while addressing the large scale of online content and even the dark side of the Internet [6, 7]. These potential threats raise a rising concern about the stability and sustainability of volunteer-based content moderation.

Algorithm-based moderation opens a new avenue for platform content moderation [8-10]. For instance, Facebook has applied advanced machine learning techniques to detect pornographic content before it can be viewed and shared [2, 6]. Compared to humans, algorithm-based moderation has high scalability to handle massive online content comprehensively and instantly.

While platforms are moving forward to the technological mode of content moderation, it may have mixed impacts on volunteer moderators' contribution to content moderation. On the one hand, volunteer moderators are both community managers and community users. When moderation tasks get automated, they may walk away from their community manager role and decrease their interaction with community users. On the other hand, delegating moderation tasks to machines can also result in an increase in volunteers' engagement. The automation of moderation tasks will offer a more efficient and healthier working environment for human moderators. The improved working environment can in turn enhance their commitment and engagement to their moderation tasks [11, 12]. Simultaneously, volunteers can also go beyond enforcing compliance in community rules by devoting more effort to advanced moderations that require more care [13].

However, relevant academic research on human-bot collaboration on content moderation is still in its infancy. First, platform governance literature has investigated human-centered community governance and identified its positive impact on members' commitment and contribution [14]. However, few studies have considered the role of algorithms in community governance. Second, human-machine collaboration literature has investigated various contexts, including the traditional business sector [15] and the online platforms [16]. However, the collaboration in the content moderation context is yet to study. Third, a growing number of content moderation literature have investigated bot usage [14, 16] in recent years. However, those work is mainly from qualitative

perspectives but provides limited quantitative empirical evidence on bot moderators' impact on human moderators' voluntary engagement [4, 17].

To fill the above research gaps, we investigate the impact of algorithm-based moderation tools (herein referred to as 'bot moderators') on human moderators' behavior. Formally, we want to approach this research objective by investigating three research questions:

RQ1: How does introducing machine-powered platform content moderation (i.e., bot moderators) affect human moderators' moderation-related effort?

RQ2: How does introducing bot moderators affect the amount of effort human moderators spent on different types of moderation (i.e., corrective and supportive moderation)?

RQ3: What community characteristics modify the impact of the introduction of bot moderation on human moderators' activity?

To answer our research questions, we select Reddit as our research context, given its highly autonomous communities and well-documented moderation records. With the rapid increase of user base on Reddit, several bot moderators have been created to facilitate routine content moderation. We collect moderation records and user and human moderator participation from 156 subreddits on Reddit from 2013 to 2014. We identify the automation of moderation tasks from the public moderation records by bot moderators. We employ an advanced natural language processing technique, BERT (Bidirectional Encoder Representations from Transformers), to identify human moderators' activities. A Difference-in-Differences model is then applied to estimate the impact of moderation automation.

Results from our econometric analyses suggest that after bot moderators are implemented, volunteer human moderators perform more moderation-related activities. Human moderators make 20.2% more corrective moderations to enforce community guidelines. Meanwhile, they also provide their community with more supportive comments — offering 14.9% more explanations. Notably, the effect manifests primarily among communities with large user bases and detailed guidelines, suggesting that community needs for moderation are the key factors driving more voluntary contributions. Overall, our results indicate that delegating content moderation tasks to algorithms augments volunteer moderators' role as community managers. The increased moderation-related engagements, especially supportive behaviors, alleviates the common concerns of de-humanization in the automation of platform content moderation.

2. Related Literature and Theories

The influence of machines on human labor has been discussed in the literature in recent years [15, 16]. Prior studies come mainly from a functional perspective to investigate the relationship between humans and machines and their collaborative outcomes. By comparing their work capabilities, machines have shown the potential to substitute humans in positions that emphasize routine and standard duties [18]. From this perspective, in the context of online content moderation, machines and algorithms can also substitute humans in achieving large-scale moderation tasks and result in a decreased need for human moderators to engage in community governance.

However, not all moderation tasks can be entirely handled by bot moderators. Compared to humans, bot moderators have limited capability to understand contexts and subjects. Some high-level community rules such as detecting satirical languages and hate speech still rely on and are subject to human interpretation. Moreover, functional substitution is not the only driver of changes in human moderators' behavior. Human moderators, particularly those moderating online communities, are voluntary labor on the platform. Abundant literature on volunteerism suggests that task-related and emotion-related support are two key factors in volunteers' engagement and commitment [11, 12, 19, 20]. Automating routine moderation tasks with bot moderators will offer human moderators a more supportive and efficient environment [12].

Bots can also provide human moderators with more emotional support, thus creating a more pleasant working environment [1]. Bots can help human moderators filter out inappropriate content first and protect moderators from harassment and harmful content [21]. A healthy and safe working environment will make volunteer work more attractive and stimulate more voluntary engagements from human moderators.

Regarding the content moderation category, by interviewing 56 volunteer moderators from Reddit, Twitch, and Facebook, Seering et al. [10] documented a list of general duties that moderators need to fulfill. Based on governance priority and effort that a moderation requires, these duties can be classified into two categories: basic and advanced moderations. Basic moderations, such as monitoring community activity and removing rule-violating content, are the top priority for human moderators because these efforts are critical to the safety and integrity of a community. In contrast, advanced moderations, such as responding to community members, seeding content, and contributing to discussions, are corresponding efforts to foster a more engaging community. It requires moderators to invest

more effort into the community nurturing in addition to enforcing the rule compliance.

Similarly, Ruckenstein and Turunen [17] summarize two major types of content moderations: logic of choice and logic of care. The logic of choice is a corrective type of governance (i.e., removing inappropriate content). In contrast, the logic of care emphasizes supportive moderation (i.e., explaining community rules and offering suggestions to community members). Both corrective and supportive moderation are critical factors for community building and growth. On the one hand, corrective moderation is an essential component for communities to maintain their integrity and order. Srinivasan et al. [22] studied a typical type of corrective moderation—content removal—and found that deleting problematic content would reduce community noncompliance rates. Supportive moderation, on the other hand, would stimulate healthier and more constructive interactions in the community. It is associated with more desired content for their community members and works as the driving force for community sustainability [17,23]. Supportive moderation complements corrective moderation in achieving more effective platform governance. For example, Jhaver et al. [24] found that while executing content removal, offering sufficient explanations would significantly reduce the possibility of users' future rule violations.

Given the importance of both corrective and supportive moderations, it is critical to investigate how bots affect either type of human moderation, especially the supportive moderation that is more desired.

3. Empirical Setting

3.1. Research Context

We collected our data from Reddit, a large online discussion and community website. It comprises more than 130,000 active communities (termed subreddits) as of 2021 covering a wide range of subjects such as world news, sports, writing, and movies. Like other dominant social media platforms such as Twitter, users can post content, leave comments, and upvote/downvote others' content on Reddit. Currently, Reddit has more than 430 million monthly active users, and it has become the fifth most visited user-generated content platform in the world as of December 2019.¹

Reddit used to solely rely on volunteers to manage all their communities. A majority of human moderators are community users who have a solid understanding of the community identity and rules. These volunteer human moderators design community rules and monitor

community activity regularly. When a rule violation happens, they are also authorized to remove content and suspend users.

Since 2012, subreddits started to utilize bot moderators to assist their platform content moderation. Two characteristics make Reddit to be an ideal research context for our study: data availability and transparency. Due to security and intellectual property concerns, most platforms do not disclose details of the technical strategies used for their content moderation. Moreover, for those platforms, we neither observe moderation records for human moderators nor bot moderators. In contrast, Reddit makes both task delegation and moderation records available to the public. On Reddit, subreddits use bots to assist with their community moderation [19, 25]. These bot moderators can automatically inspect newly submitted content and conduct a series of actions accordingly. Each bot moderator has its profile page, which lists its creation date and historical moderation records. Most importantly, all records are presented in the form of public comments. These comments document details such as moderation time, action, explanations, and suggestions to the content creator. Human moderators perform their moderations in a similar manner. When they find a rule violation, they can perform moderation and inform the content creators through public comments, submission flairs, and private messages. Among these three communication channels, public comment is the dominant approach [5]. Public moderation records are the primary data source for our study. And based on these records, we recover the moderation automation timeline and investigate its impacts on human moderators' voluntary activities.

3.2. Data Collection and Measures

We choose 'AutoModerator' as the focal bot in our study. AutoModerator was introduced in 2012, and it has become the most influential bot moderator on Reddit. The bot generates thousands of moderation records every day. AutoModerator is a moderation tool with various functions. To delegate tasks to AutoModerator, human moderators need only follow simple syntax and add bot-related codes into their community meta page. The typical moderation tasks that AutoModerator can achieve include restricting users under a certain level of reputation on the platform and preventing users from posting content from sources that are blacklisted. Subreddits on the Reddit platform can integrate the AutoModerator functions to deploy it as a bot moderator based on their community guidelines and moderation needs.

¹ <https://redditblog.com/2019/12/04/reddits-2019-year-in-review/>

Our data collection includes four steps. First, we collect all moderation records by AutoModerator from 2013 to 2014 using PushShift API.² We choose this observation period because AutoModerator’s function was relatively stable after being released for a year. Further, more subreddits started incorporating it into their moderation since 2013. Thus, the selection of subreddits in that period would contain ample variation in the corresponding empirical analyses. Second, we use the time when the first AutoModerator’s moderation was performed as the proxy adoption date of AutoModerator in a subreddit. With the complete moderation records collected in the first step, we observe that 156 subreddits adopted AutoModerator from 2013 to 2014. Next, we identify human moderators for each of the studied subreddits. Reddit grants moderators a “mod” flair to signal their role as community managers, and this flair is shown along with moderators’ comments. Taking advantage of this platform design feature, we obtain the list of users with this flair and further collect their comments during the observation window. Lastly, with AutoModerator’s and human moderators’ comments, we perform machine learning to extract the automated rules and human moderators’ engagement types.

Table 1. Examples of different types of moderator comments

Category	Example
Policing	“Sorry, this is a repost within 3 months. Therefore, I am removing it.”
Explanation	“We made this decision because you did not include the resolution in the title and because of man-made structures and you did not include the location in the title.”
Suggestion	“There are instructions in the FAQ as well as our other submission guidelines. This would be more appropriate in r/villageporn. Feel free to resubmit with the resolution in the title. Thanks!”
Casual talk	“Thank you and it is even more beautiful in the reality especially the very special light of Iceland”

We performed two aspects of data pre-processing. First, on the AutoModerator side, we extract automated moderation tasks and task implementation time to recover the timeline of moderation delegation. We utilize this timeline to construct the independent variable of interests (i.e., *AutoMatedTasks_{it}* and *After_{it}*) in our following analyses. Given that AutoModerator

mostly follows limited formats to frame a moderation record, to achieve this goal, we can employ a rule-based approach to extract the automated task and its implementation time.

Second, on the human moderator side, our pre-processing task is to identify their activities and moderation types. We seek to perform a theory-driven annotation (labeling) task for a sample of human moderators’ comments for four categories: policing, explanation, suggestions, and casual talk. Table 1 illustrates examples for each category. Among these four categories, policing, explanation, and suggestion belong to the moderation-related comments, whereas casual talk represents the informal interaction that human moderators have with community users. With regards to the moderation type, policing represents corrective moderation, whereas suggestions and explanations are supportive moderations.³ Compared to processing of bot comments, this task is relatively challenging due to the diversity and complexity of human language. Our methods need to go beyond the vocabularies to grasp the semantic meaning and purpose of a comment.

Specifically, we utilize BERT, a deep learning approach, to classify human moderators’ moderation types [26]. BERT was proposed in 2018, and it has quickly become the state-of-the-art natural language processing (NLP) technics and has achieved exceptional performance in NLP tasks such as classification, Q&A, and commonsense reasoning. Compared to 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 respective position. Therefore, BERT can capture the meaning of comments regardless of the vocabulary choice and sentence sequence. We use BERT-based classifiers to identify moderation-related comments as well as detailed moderation types.

Operationally, similar to other deep learning-based NLP techniques, we use the pre-trained BERT model and then fine-tune the parameters with a set of labeled data. Pre-trained BERT can capture the context-free meaning of the comments. The labeled data will provide more details about our context. We obtain the pre-trained model from Google Research, and this model is trained on large corpora collected from Wikipedia and book chapters.⁴ Then, we randomly sampled 1,017 comments from our whole dataset and manually labeled these comments for each category. Two graduate students in the computer science master’s program at a

² <https://pushshift.io/>

³ Following real-world practice, a comment can only be either moderation-related or not, but a moderation-related comment can have both corrective and supportive components. Thus, in our annotation

task, we restrict casual talk as exclusive from three moderation-related categories, and we allow the three moderation types to coexist in a human moderator’s comment.

⁴ <https://github.com/google-research/bert>

large public research university labeled all comments. They achieved 85% consistency in their labeling, for the remaining 15% comments, they further discussed their discrepancies in annotation and reached agreement as well. We further feed the pre-trained model with the labeled data to fine-tune the parameters to better work for this task and context. For each comment category, we create a classifier, for a total of four classifiers. Lastly, we use the fine-tuned models to predict all collected human comments. We ran the whole classification process on the Google Cloud Platform and used TPUs to perform all BERT-related processing.

Table 2. Performance of BERT-based classifiers

Category	Accuracy	Precision	Recall	F1 Score
Policing	97.34%	96.99%	95.68%	96.33%
Explanation	95.37%	95.02%	95.22%	95.12%
Suggestion	95.96%	96.84%	95.15%	95.99%
Casual talk	92.52%	94.47%	84.66%	89.30%

Table 2 shows the classification results of the four classifiers. We can see that, on average, we have achieved above 90% on the F1 score in four classification metrics. Notably, the policing classifier has reached 97.34% accuracy, 96.99% precision, 95.68% recall, and 96.33% F1 score. Compared to typical NLP classification tasks, our BERT-based classifiers have achieved a satisfying prediction performance.

Lastly, with all collected data, we construct a series of variables for the corresponding data analyses. We aggregate our data at the subreddit-monthly level. We used the number of labeled comments created by human moderators to measure their various types of engagement. In addition to the data mentioned above, we also collect regular users' comments from each subreddit and use it as the proxy for user engagement in that community. Table 3 depicts detailed descriptions and summary statistics of all measures in this study.

4. Empirical Analyses and Results

4.1. Identification Strategy and Main Model

Considering the automation of moderation tasks is a shock to the community and that communities adopted bots for moderation at different times, we apply Difference-in-Differences (DiD) as our identification strategy. Specifically, the treatment group comprises

communities that automated moderation tasks, whereas the control group includes communities that relied entirely on volunteer human moderators at each observational point. The same approach is applied in prior literature studying the impacts of ride-sharing entry [27, 28] and technology adoption [29]. By comparing human moderators' activities of the treatment group and the control group, we are able to estimate how the integration of bot moderators to the platform content moderation affects volunteer human moderators' engagement.

Table 3. Variables and descriptive statistics

Variable	Description	Mean	S.D.
AutoMate dTasks _{it}	It represents the number of moderation tasks delegated to AutoModerator on subreddit <i>i</i> in month <i>t</i> .	2.231	3.342
Num_Us er_Cmts _{it}	It represents the number of comments that users created on subreddit <i>i</i> in month <i>t</i> .	89,351.66	358182.9
Num_Mo d_Partici pation _{it}	It represents the number of comments that human moderators created on subreddit <i>i</i> in month <i>t</i> .	322.119	668.175
Num_Mo d_Casual _{it}	It represents the number of casual comments that human moderators created on subreddit <i>i</i> in month <i>t</i> .	215.862	446.524
Num_Mo d_Moder ation _{it}	It represents the number of moderation-related comments that human moderators created on subreddit <i>i</i> in month <i>t</i> .	106.257	318.407
Num_Mo d_Policin g _{it}	It represents the number of policing type of comments that human moderators created on subreddit <i>i</i> in month <i>t</i> .	63.882	247.128
Num_Mo d_Explan ation _{it}	It represents the number of explanation type of comments that human moderators created on subreddit <i>i</i> in month <i>t</i> .	80.166	282.480
Num_Mo d_Sugges tion _{it}	It represents the number of suggestion type of comments that human moderators created on subreddit <i>i</i> in month <i>t</i> .	84.352	256.130
Num_Mo ds _{it}	It represents the number of active human moderators on subreddit <i>i</i> in month <i>t</i> .	3.827	4.440

Formally, our empirical model is presented in Model (1). *ModParticipations_{it}* denotes the dependent variable, volunteer moderators' activities on the focal community. *After_{it}* is the key independent variable. It is a binary variable, and the value of 1 means the moderation comment was generated after the subreddit implemented AutoModerator. Additionally, each human moderators' activity may be subject to the

changes in the number of active human moderators and the number of submissions that their community received. Thus, we add two control variables—the number of active human moderators (i.e., Num_Mods_{it}) and the number of user comments (i.e., $Num_User_Cmts_{it}$) in respective subreddits—to the analysis. Another source of empirical concern is the unobserved intrinsic difference among communities and temporary shocks on the platforms. Therefore, we add the subreddit fixed effect θ_i and monthly fixed effect δ_t to control the unvarying subreddit and time impacts. ε_{it} is the error term.

$$ModParticipations_{it} = \alpha + \beta \times After_{it} + \gamma \times Controls_{it} + \delta_t + \theta_i + \varepsilon_{it} \quad (1)$$

4.2. Main Results

Table 4 shows several interesting findings. First, we can see that the automation of platform content moderation encourages human moderators to perform more corrective behavior. Their policing activities increase by 20.2% after their community delegated moderation tasks to AutoModerator. Second, more supportive moderator behaviors also emerged. The explanation type of comments increases by 14.9%. The estimates of our control variables also align with real-world experience. When there are more user submissions and more active moderation team members, moderators' activities will also increase accordingly. These results reduce the concerns that our empirical analysis model may suffer from multicollinearity.

Overall, these results indicate that automating content moderation is positively associated with policing-related engagement in the community. Meanwhile, moderators also provide more support to their community members by offering more explanations about their community policies.

5. Empirical Extension

5.1. Moderating Effect of Community Size

We perform two heterogeneity analyses to examine how community characteristics moderate the effects of bot moderators on human moderators' activities to further uncover some underlying mechanisms. The first community characteristic that we investigate is community size [30, 22]. Community size may play a role in the moderators' voluntary engagement for two reasons. First, community size implies community popularity and the resources (e.g., community content)

available to their community members [30]. Moderators in popular communities receive more attention and higher prestige, as their moderation generates a more considerable impact. Such status and exposure will result in higher levels of commitment to community and engagement.

Second, compared to subreddits with fewer members, large communities usually have more urgent moderation needs to accommodate the massive amounts of submissions and diverse community requirements. Volunteer moderators in popular communities are expected to meet growing numbers of user requests and deal with a large corpus of user-generated content. When algorithm-based moderation tools become available, the augmentation impact will be more prominent for moderators from popular communities.

Table 4. The impact of AutoModerator on users' and human moderators' participation

Variables	Moderator Role			User Role
	Num_Mod_Policing	Num_Mod_Expla_nation	Num_Mod_Sugges_tion	Num_Mod_Casual_Talk
After	0.202** (0.0909)	0.149* (0.0838)	0.0745 (0.0702)	-0.0570 (0.0638)
Num_User_Cmts	0.110*** (0.0364)	0.119*** (0.0374)	0.108*** (0.0353)	0.175*** (0.0370)
Num_Mods	1.442*** (0.0890)	1.593*** (0.0824)	1.727*** (0.0724)	1.911*** (0.0746)
Constant	-1.061*** (0.307)	-0.985*** (0.318)	-0.781** (0.301)	-0.602* (0.325)
# of subreddits	156	156	156	156
N	3,744	3,744	3,744	3,744
R-squared	0.670	0.604	0.555	0.604

Note: (1) Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; (2) For all dependent variable and the number of commenters, we use log transformation (i.e., log(x+1)) to accommodate skewed distribution and zeros.; (3) Subreddit, month and year fixed effects are included.

To test the moderating impacts 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 and construct a binary variable $PopSubreddit_i$ to denote subreddits with relatively more members. We add this term and an interaction term between $PopSubreddit_i$ and $After_{it}$ into Model (1). Table 5 presents the estimation results. With the estimation of the automation decision (i.e., $After_{it}$)

and the interaction term (i.e., *After* × *PopSubreddi*), we note that automating moderation tasks does not change moderators' policing and explanation activities in small communities but positively affect voluntary moderation in larger communities. Moreover, we also find that usage of bots oppositely affects moderators' suggestion activity in popular and non-popular subreddits. There was a 14.9% decrease in moderator suggestions on the small subreddits, and a 29.4% increase on the popular communities. With regard to the informal interaction with community members, we see a 16.2% decrease in both large and small communities. This empirical evidence suggests that machines weaken moderators' role as regular community members regardless of community popularity.

Table 5. The moderating effect of community size

Variables	Moderator Role			User Role
	Num_Mod_Policing	Num_Mod_Expla_nation	Num_Mod_Sugges_tion	Num_Mod_Casual_Talk
After	-0.0690 (0.0930)	-0.122 (0.0941)	-0.149* (0.0889)	-0.162* (0.0882)
After × PopSubreddit	0.536*** (0.167)	0.537*** (0.152)	0.442*** (0.130)	0.207 (0.129)
Num_Us_er_Cmts	0.114*** (0.0342)	0.123*** (0.0353)	0.112*** (0.0331)	0.177*** (0.0362)
Num_Mo_ds	1.428*** (0.0902)	1.579*** (0.0836)	1.716*** (0.0733)	1.906*** (0.0744)
Constant	-1.073*** (0.289)	0.997*** (0.300)	0.791*** (0.280)	-0.606* (0.318)
# of subreddits	156	156	156	156
N	3,744	3,744	3,744	3,744
R-squared	0.419	0.478	0.529	0.606

Note: (1) Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; (2) For all dependent variable and the number of commenters, we use log transformation (i.e., log(x+1)) to accommodate skewed distribution and zeros.; (3) Subreddit, month and year fixed effects are included.

Overall, our results suggest that the human and moderation relationship is subject to the community size. On small communities, bots substitute volunteer moderators as evidenced by the decreased moderation-related and informal interaction. In contrast, bots exhibit positive impacts on moderators from large communities by augmenting their role as community managers. Such differences may be caused by the attractiveness of

communities and the influence of moderators. When a community has a larger user base and more engaging content, moderators have greater impact and prestige on the platform. They also face more extensive needs in moderating community content appropriately and promptly. Therefore, algorithm assistance enables these volunteers to stick with their moderator role and even achieve better performance, rather than leading them to disengage from the community.

5.2. Moderating Effect of Scope of Work

We next perform the second heterogeneity test on scope of work. Similar to rules and culture in organizations, community guidelines specify its identity and norms. These guidelines provide quick guidance for users, especially newcomers, about the community focus and participation requirements. Meanwhile, they also specify volunteer moderators' scope of work and serve as a reference for them while governing their community.

We consider how scope of work will moderate the impacts of bots on human moderators' behavior for two reasons. First, when a community specifies a broad scope of work in content moderation, moderators will have less ambiguity in their role and better understand their responsibilities. Prior literature in the organizational context has examined that role clarity positively affects the employee's job commitment and satisfaction [31-33]. We predict that when algorithms automate some routine moderation tasks, moderators from subreddits with more community guidelines are more likely to maintain their continued efforts and take care of other moderation tasks that cannot be automated by bot moderators. Second, clear and comprehensive scope of work officially document more specific circumstances that volunteers may run into in their moderation and the corresponding solutions. It provides volunteers more organizational support to perform moderations and justify their actions. Thus, it becomes easier for moderators to conduct advanced community moderation, such as explaining their moderation decisions and recommending other available resources. Emotionally, detailed and thorough rules will also protect volunteers from having unnecessary conflicts with their members due to unclear and subjective moderation.

To test the moderating effect of a scope of work, we search all studied subreddits' historical pages during our

observational period.⁵ Then, we scrape the public community rules on these pages. We measure scope of work using the number of rules posted on a community. The more community rules, the broader a scope of work. We split the studied subreddits into two groups using the median value of the number of community rules and then label subreddits with more rules as group with a broad scope of work (i.e., *BroadSOW_i*). Next, we add the new term, *BroadSOW_i*, and an interaction term between *BroadSOW_i* and *After_{it}* into Model (1). See Table 6 for the results. The estimation is consistent with what we expected. Bots only show positive impacts on volunteers' moderation engagement in subreddits with more moderation tasks. In contrast, communities with fewer rules experience the same or even decreased volunteer effort in managing their communities after delegating tasks to bots.

Table 6. The moderating effect of scope of work

Variables	Moderator Role			User Role
	Num_Mod_Policing	Num_Mod_Expla_nation	Num_Mod_Sugges_tion	Num_Mod_Casual_Talk
After	-0.0735 (0.112)	-0.138 (0.104)	-0.174* (0.0910)	-0.128 (0.0790)
After × BroadSOW	0.557*** (0.170)	0.580*** (0.154)	0.504*** (0.128)	0.144 (0.127)
Num_Us_er_Cmts	0.108*** (0.0327)	0.116*** (0.0337)	0.106*** (0.0322)	0.175*** (0.0366)
Num_Mods	1.406*** (0.0874)	1.556*** (0.0806)	1.694*** (0.0703)	1.902*** (0.0736)
Constant	-1.005*** (0.276)	- (0.286)	- (0.275)	-0.587* (0.322)
# of subreddits	156	156	156	156
N	3,744	3,744	3,744	3,744
R-squared	0.419	0.480	0.531	0.605

Note: (1) Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1; (2) For all dependent variable and the number of commenters, we use log transformation (i.e., log(x+1)) to accommodate skewed distribution and zeros; (3) Subreddit, month and year fixed effects are included.

6. General Discussion

In recent years, digital platforms have started widely adopting algorithms to offload human moderators from ever increasing burden of routine moderation tasks. As platforms are moving to the technical mode of platform content moderation with help from algorithms, researchers have raised a concerns on whether algorithms would lead volunteer human moderators to reduce their community contributions. In this study, we collect moderation records from Reddit and investigate the impact of machine-powered content moderation on volunteer human moderation. With data collected from 156 subreddits, we found that delegating moderation to machines augments volunteer moderators' role as community managers. Human moderators present more moderation-related engagement, including both corrective and supportive interactions with their community members. Notably, our results indicate that such effects manifest among communities with large user bases and detailed community guidelines, suggesting that community needs for moderation is the driving factor for volunteer moderators' increased contributions.

Our results first dispel concerns from community managers that applying machines to platform regulation will drive volunteer moderators away from their community and cause less community engagement. In contrast, equipping volunteer moderators with more technological moderation tools will augment their role as community managers. Machines empower these volunteer moderators to perform more comprehensive community moderation, which may require subjective interpretation and a better understanding of the context. Most importantly, while achieving more comprehensive corrective engagement, volunteer moderators contribute more supportive moderation to their community. This result resolves concerns of the de-humanizing effects of algorithmic regulation as machines will stimulate more positive engagement from volunteer moderators. Our finding also imply that compared to content moderation models that solely rely on human moderation, human-machine collaboration brings more sustainability to the volunteer moderation team and in turn helps the community grow.

Our study suggests that not all communities would benefit from moderation automation. Popular communities with a massive user base will enjoy the increase in volunteer moderators' contributions. However, for small communities with limited influence

⁵ We use <https://archive.org/web/> in our study to find out historical subreddit pages.

and attractiveness, platform practitioners can encourage moderators to establish clear and detailed community guidelines. By enlarging and specifying moderators' scope of work with more detailed community guideline, automation can also stimulate more voluntary contributions in the community.

Our work further suggests that platform practitioners need to be careful about the decreasing informal conversations between moderators and community members when regulation is automated. Interpersonal communication is a critical factor in enhancing bond-based attachment [34]. Ideal platform content moderation also needs to balance formal regulation and informal social control [35]. Our results indicate that communities with relatively small user bases and non-entertainment related communities have shown a more significant decrease. Considering the importance of interpersonal interaction to communities, platform practitioners should evaluate the impact of reduced informal engagement and guide volunteer moderators to avoid potential adverse outcomes.

We understand that there are some drawbacks to our research. First, we currently measure the volunteer moderators' scope of work by the number of rules in their community. However, the scope of work each rule covers might be different from each other. In the future, we can consider applying alternative measures for the scope of work, such as whether the rules require subjective judgment or pertain to multiple scenarios.

Second, our study only investigates the human-machine collaboration from the angles of changes in volunteer human moderators' community engagement. We did not consider the new effort and time that human moderation spent creating and maintaining this collaboration (i.e., human-in-the-loop). For example, to assimilate bots into community moderation, the moderation teams on Reddit need to have some tech-savvy members to design bots, fix algorithmic errors, and ensure the bots are aligned with updated community guidelines [4]. However, these costs occur in the background, and are difficult for researchers to observe. In this study, we only focus on the changes in volunteers' moderation, the direct interaction with community members. Nevertheless, human-in-the-loop would be an intriguing topic for future research. It will generate valuable insights into moderation teams' work design and management.

Third, while 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 in integrating algorithms into daily content moderation. Therefore, it is necessary to go beyond our research context and examine the external validity on other platforms. Achieving this goal requires understanding business

logic for different platforms and effective approaches to access data. Therefore, we will open this avenue to future researchers and look forward to more insights into good human-machine regulation on other platforms.

7. References

- [1] J. N. Matias, "The Civic Labor of Volunteer Moderators Online", *Social Media+ Society*, 2019, 5(2), pp. 2056305119836778.
- [2] S. T. Roberts, "Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation", PhD diss., University of Illinois at Urbana-Champaign, 2014.
- [3] L. Zheng, C. M. Albano, N. M. Vora, F. Mai, and J. V. Nickerson, "The Roles Bots Play in Wikipedia", *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3.
- [4] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Human-machine Collaboration for Content Regulation: The Case of Reddit Automoderator", *ACM Transactions on Computer-Human Interaction*, 2019, 26(5), pp. 1-35.
- [5] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, "Did You Suspect the Post Would be Removed? Understanding User Reactions to Content Removals on Reddit", *Proceedings of the ACM on Human-computer Interaction*, 2019, 3.
- [6] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*, Yale University Press, 2018.
- [7] J. N. Matias, "Preventing Harassment and Increasing Group Participation through Social Norms in 2,190 Online Science Discussions", *Proceedings of the National Academy of Sciences*, 2019, 116(20), pp. 9785-9789.
- [8] K. Gollatz, F. Beer, and C. Katzenbach, "The Turn to Artificial Intelligence in Governing Communication Online", 2018.
- [9] H. L. Hammer, "Automatic Detection of Hateful Comments in Online Discussion", In *International Conference on Industrial Networks and Intelligent Systems*, 2016, pp. 164-173.
- [10] J. Seering, T. Wang, J. Yoon, and G. Kaufman, "Moderator Engagement and Community Development in the Age of Algorithms", *New Media & Society*, 2019, 21(7), pp. 1417-1443.
- [11] K. Alfes, A. Shantz, and C. Bailey, "Enhancing Volunteer Engagement to Achieve Desirable Outcomes: What Can Non-profit Employers Do?" *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 2016, 27(2), pp. 595-617.
- [12] D. H. Smith, "Determinants of Voluntary Association Participation and Volunteering: A Literature Review", *Nonprofit and Voluntary Sector Quarterly*, 1994, 23(3), pp. 243-263.
- [13] N. Karusala, A. Vishwanath, A. Kumar, A. Mangal, and N. Kumar, "Care as a Resource in Underserved Learning Environments", *Proceedings of the ACM on Human-Computer Interaction*, 2017, 1.

- [14] Y. Ren, and R. E. Kraut, "Agent-based Modeling to Inform Online Community Design: Impact of Topical Breadth, Message Volume, and Discussion Moderation on Member Commitment and Contribution", *Human-Computer Interaction*, 2014, 29(4), pp. 351-389.
- [15] J. Dixon, B. Hong, and L. Wu, "The Employment Consequences of Robots: Firm-Level Evidence", Available at SSRN 3422581, 2019.
- [16] X. Luo, S. Tong, Z. Fang, and Z. Qu, "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases", *Marketing Science*, 2019, 38(6), pp. 937-947.
- [17] M. Ruckenstein, and L. L. M. Turunen, "Re-humanizing the Platform: Content Moderators and the Logic of Care", *New Media & Society*, 2020, 22(6), pp.1026-1042.
- [18] D. H. Autor, and D. Dorn, "The Growth of Low-skill Service Jobs and the Polarization of the US Labor Market", *American Economic Review*, 2013, 103(5), pp.1553-97.
- [19] H. Bang, S. Ross, and T. G. Reio, "From Motivation to Organizational Commitment of Volunteers in Non-profit Sport Organizations", *Journal of Management Development*, 2013, 32(1), pp. 96-112.
- [20] M. L. Vecina, F. Chacón, D. Marzana, and E. Marta, "Volunteer Engagement and Organizational Commitment in Nonprofit Organizations: What Makes Volunteers Remain within Organizations and Feel Happy?" *Journal of Community Psychology*, 2013, 41(3), pp. 291-302.
- [21] S. Jhaver, P. Vora, and A. Bruckman, "Designing for Civil Conversations: Lessons Learned from ChangeMyView", *Georgia Institute of Technology*, 2017.
- [22] K. B. Srinivasan, C. Danescu-Niculescu-Mizil, L. Lee, and C. Tan, "Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community", *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3, pp. 1-21.
- [23] B. Yu, K. Spiel, and L. Watts. "Supporting Care as a Layer of Concern: Nurturing Attitudes in Online Community Moderation.", In *Sociotechnical Systems of Care: A CSCW18 Workshop*, 2018.
- [24] S. Jhaver, A. Bruckman, and E. Gilbert, "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit", *Proceedings of the ACM on Human-Computer Interaction*, 2019, 3, pp. 1-27.
- [25] C. Fiesler, J. Jiang, J. McCann, K. Frye, and J. Brubaker, "Reddit Rules! Characterizing an Ecosystem of Governance", *Proceedings of the International AAAI Conference on Web and Social Media*, 2018, 12(1). 2018.
- [26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018, arXiv preprint arXiv:1810.04805.
- [27] G. Zervas, D. Proserpio, and J. W. Byers, "The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry", *Journal of Marketing Research*, 2017, 54(5), pp. 687-705.
- [28] G. Burtch, S. Carnahan, and B. N. Greenwood, "Can You Gig it? An Empirical Examination of the Gig Economy and Entrepreneurial Activity", *Management Science*, 2018, 64(12), pp. 5497-5520.
- [29] T. F. Tan, and S. Netessine, "At Your Service on the Table: Impact of Tabletop Technology on Restaurant Performance", *Management Science*, 2020, 66(10), pp. 4496-4515.
- [30] B. S. Butler, P. J. Bateman, P. H. Gray, and E. I. Diamant, "An Attraction-selection-attrition Theory of Online Community Size and Resilience", *MIS Quarterly*, 2014, 38(3), pp. 699-729.
- [31] T. F. Lyons, "Role Clarity, Need for Clarity, Satisfaction, Tension, and Withdrawal", *Organizational Behavior and Human Performance*, 1971, 6(1), pp. 99-110.
- [32] J. H. Donnelly, and J. M. Ivancevich, "Role Clarity and the Salesman: An Empirical Study Reveals that Perceived Role Clarity May be an Important Factor in Maximizing a Salesman's Job Performance", *Journal of Marketing*, 1975, 39(1), pp. 71-74.
- [33] S. Hassan, "The Importance of Role Clarification in Workgroups: Effects on Perceived Role Clarity, Work Satisfaction, and Turnover Rates", *Public Administration Review*, 2013, 73(5), pp. 716-725.
- [34] Y. Ren, F. M. Harper, S. Drenner, L. Terveen, S. Kiesler, J. Riedl, and R. E. Kraut, "Building Member Attachment in Online cMmunities: Applying Theories of Group Identity and Interpersonal Bonds", *MIS Quarterly*, 2012, 36(3), pp. 841-864.
- [35] M. Williams, "Policing and Cybersociety: The Maturation of Regulation Within an Online Community", *Policing & Society*, 2007, 17(1), pp. 59-82.