How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment

Leijie Wang
wanglj17@mails.tsinghua.edu.cn
Tsinghua University
Beijing, China

Haiyi Zhu
haiyiz@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Machine learning-based predictive systems are increasingly used to assist online groups and communities in various content moderation tasks. However, there are limited quantitative understandings of whether and how different groups and communities use such predictive systems differently according to their community characteristics. In this research, we conducted a field evaluation of how content moderation systems are used in 17 Wikipedia language communities. We found that 1) larger communities tend to use predictive systems to identify the most damaging edits, while smaller communities tend to use them to identify any edit that could be damaging; 2) predictive systems are used less in content areas where there are more local editing activities; 3) predictive systems have mixed effects on reducing disparate treatment between anonymous and registered editors across communities of different characteristics. Finally, we discuss the theoretical and practical implications for future human-centered moderation algorithms.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; *Empirical studies in collaborative and social computing*;

KEYWORDS

Content moderation, Wikipedia, Causal inference, Fairness, Online communities

ACM Reference Format:
Leijie Wang and Haiyi Zhu. 2022. How are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3531146.3533147

1 INTRODUCTION

Content moderation is a common issue in almost every online space that allows users to generate content. A 2017 Pew survey found that four in ten Americans had personally experienced online harassment [51]. Machine learning based predictive systems are widely used to moderate undesirable content in online communities [10, 23, 56, 57]. For example, Twitter has adopted anti-harassment algorithms to identify block-worthy accounts [39]. YouTube and Facebook have deployed AI-based algorithms to quickly identify posts that violate community norms for manual review [4, 27].

While algorithms have the potential to make moderation work more reliable and efficient [14], a considerable body of work has pointed out that moderation algorithms are sometimes in conflict with important values of people and communities who use them. For example, moderators believe in the importance of retaining themselves as the final judgement and therefore resist fully-automated moderation systems [61]. Researchers have also criticized that the moderation algorithms tend to prioritize the moderation efficiency but neglect other important moderation goals, including fair treatment towards anonymous editors and newcomers [32, 35].

However, previous research often relied on *qualitative* methods to investigate the conflict between moderation algorithms and human values, but relatively little work has *quantitatively* analyzed this conflict at scale. The most relevant work is Teblunthuis et al's, which measured the influence of moderation algorithms on human moderators at scale [62]. However, they did not investigate how the influence of algorithms might vary across different kinds of communities or show the interaction between community characteristics and moderation algorithms.

In this research, we study how the ORES-powered RCFilters (Recent Change Filters), a machine-assisted content moderation system, was used in 17 Wikipedia language communities. To address the increasing challenge of vandalism, Wikipedia has deployed the RCFilters to flag potentially damaging edits for manual review. The RCFilters is supported by ORES, an ML-based algorithmic scoring service that predicts the likelihood an edit is damaging [30]. The RCFilters then compares the prediction with a pre-selected threshold to determine whether that edit should be flagged or not.

We aim to understand the following two research questions in Wikipedia communities of different characteristics. In particular, we investigate two community characteristics that might shape the use of the RCFilters: **the community size** (i.e. how many new edits are made every month in a community), **the local editing activity** (i.e. how many active editors in a community are keeping an eye on an edit).

- How does this moderation system impact human moderation decisions?
- Does this moderation system help alleviate the disparate treatment between anonymous and registered editors?

This work is licensed under a Creative Commons Attribution-Share Alike International 4.0 License.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533147

To answer these questions, we use the Regression Discontinuity Design (RDD) to quantitatively evaluate the causal effect of this machine-assisted moderation system. The intuition behind the RDD is that data points immediately below and above an arbitrary cutoff are equal in expectation; comparing the outcome of the data points immediately above to the counterfactual outcome of those immediately below will hence deliver the local treatment effect. In the context of the RCFilters, the arbitrary cutoff is *the damaging threshold*. By comparing the reverting rates of edits immediately above versus immediately below the damaging threshold, we are able to measure the impact of the RCFilters on human moderation decisions and understand how the system interacts with other community characteristics at play.

We analyzed 4,000,000 edits from 17 Wikipedia language communities between 2019 and 2020 and found that 1) larger communities tend to use the RCFilters to identify the most damaging edits, while smaller communities tend to use the RCFilters to identify any edit that could be damaging; 2) The RCFilters is used less in content areas where there are more local editing activities; 3) The RCFilters has mixed effects on reducing disparate treatment between anonymous and registered editors in communities of different characteristics.

In this paper, we first review literature about content moderation in online communities and fairness of moderation algorithms. We then explain how the community size and the local editing activity influence the use of moderation algorithms in Wikipedia and propose hypotheses accordingly. In the main part, we quantify the effect of the moderation system on moderation decisions and the disparate treatment between editor groups in different types of communities. Finally, we discuss the limitations of our research, and present theoretical and design implications for future research.

2 RELATED WORK

2.1 Content Moderation in Online Communities

Despite the indispensable role online communities (e.g. Facebook, Wikipedia, and YouTube) play in our daily life, we still see with disappointment an overwhelming number of examples of spam, misinformation, and off-topic postings in our online experience. A nationally representative Pew Research Center study in 2017 found that roughly four in ten Americans had personally experienced online harassment [51]. While platforms like NPR [18] and IMDB [2] are forced to close their comment sections due to their incapability to fight against vandalism, other platforms have adopted a variety of approaches to regulate online behavior to maintain positive and productive conversations [45].

These moderation approaches can be further divided into three categories: manual moderation, fully automated moderation algorithms, and machine-assisted moderation systems. Human moderators are employed by online platforms to regulate content [1, 58] but face great challenges as the amount of content to be reviewed grows. Therefore, online communities like Facebook [4], YouTube [27], and Reddit [8] increasingly adopt algorithms to automatically identify and remove damaging content at scale. Despite the efficiency of automated algorithms to moderate content [13, 14, 23], people are concerned about their incompetency to make context-sensitive moderation decisions and potential violations against community norms [29, 61, 63]. Therefore, machine-assisted moderation systems are increasingly used.

In reality, large online platforms have established a multi-layer immune system against vandalism. In Wikipedia, both fully autonomous robots and machine-assisted systems are important components of moderation teams [24, 26, 31, 35]. Fully autonomous robots are responsible for reverting the most blatant offenders, while machine-assisted systems enable people to review more subtle damages and norm violations by filtering or flagging edits. In this paper, we study the tool, RCFilters, which flags possibly damaging content as a built-in browser widget accessible to all Wikipedia moderators.

2.2 Fairness of Content Moderation Algorithms

As algorithmic systems are widely used in both the public and private sectors for making decisions with real consequences on people's lives, there has been an increasing focus in the research community on investigating and improving the fairness of algorithmic decision-making systems [11, 67]. Fairness-aware machine learning research has abstracted fairness notions from the real-world context (e.g. equalized odds [36], demographic parity, individual fairness [16, 41]) and developed fair algorithms accordingly [12, 48]. For example, demographic parity entails a fair algorithm has parity of positive classification rates across a fixed number of protected groups, while individual fairness in [16] requires treating similar individuals similarly. On the other hand, HCI researchers have also involved human efforts in designing fair algorithms and identifying potential bias in algorithms. For example, users are empowered with interactive interfaces to visually examine the behavior of machine learning models and identify potential biases [3, 68].

Content moderation algorithms decide which post to remove and have a direct impact on our online experience. Over the past few years, there has been a heated debate over whether content moderation algorithms increase fairness or not [6, 17, 28, 42]. Prior research argued that algorithms might inherit the biases of either system developers or training datasets and then amplify them on platforms across a potentially massive scale. Binns et al. trained moderation algorithms on datasets labeled by different demographic subsets and observed that algorithm performances in the field were greatly influenced by who labeled the dataset [5]. Matamoros-Fernández argued that Twitter and YouTube embody a form of racism in their vague policies and design of moderation algorithms [52]. On the other hand, community members also consider algorithms more objective and fairer than human moderators because algorithms have some predefined, fixed criteria [39]. Therefore, it remains to be explored whether moderation algorithms increase fairness or not. In this research, we investigate the effect of a machine-assisted moderation system in Wikipedia on the disparate treatment between anonymous and registered editors in different types of communities.

2.3 Impacts of Moderation Algorithms on Content Removal

There is a growing body of work in the research community emphasizing the importance of understanding how human moderators interact with moderation algorithms in the real world. Kitchin argued that the effect of algorithms unfolds in contingent and relational ways, producing localized and situated outcomes, so empirical observations about how people engage with algorithms in practical contexts are crucial to understanding algorithms [46]. Similarly, Selbst et al. stated that it is not enough to understand algorithms in terms of their performances on the test dataset but we should also account for how real-world stakeholders respond to algorithm recommendations [53, 60].

A considerable amount of literature provides qualitative analysis of the interaction between moderation algorithms and human moderators [21, 22, 30, 69]. For example, Jhaver et al. revealed the challenges and benefits of human moderators' using automated tools in the sociotechnical context of Reddit [38]. Geiger et al. analyzed how moderation systems transform the complicated moderation work into mundane affairs performed by human moderators [25].

In contrast, there is relatively little work that quantitatively measures the impact of moderation algorithms on manual moderation decisions in the real world. Geiger et al. took advantage of breakdowns of an automated algorithm in Wikipedia and revealed that the algorithm accelerates the removal of undesirable content [23]. Compared to automated algorithms, it is even more difficult to measure the effect of machine-assisted moderation systems because of the mixture of individual discretion and algorithm recommendations inherent in this kind of systems

Recently, Teblunthuis et al. have managed to demonstrate quantitatively that the RCFilters, a machine-assisted algorithmic system deployed in Wikipedia, reduces the bias against newcomer and anonymous editors in moderation work [62]. However, it only captures parts of the complexity of decision-making scenarios on online platforms [15, 33]. For example, online communities have moderation goals that vary according to their community size and therefore might exploit this system differently. We further attempt to understand this algorithmic system in Wikipedia's sociotechnical context. We borrow methods from the causal analysis to quantitatively estimate its influence and further understand the role this algorithm plays in moderation work. Our work differs from that of Teblunthuis et al. in that we aim to explore how the influence of the system on moderation work and disparate treatment between editor groups might vary in different types of Wikipedia communities, while they treated the Wikipedia community as a whole and inquired whether the algorithmic treatment towards editor groups complies with community norms in general.

3 STUDY CONTEXT: WIKIPEDIA MODERATION

Guided by the principle that *anyone can edit*, Wikipedia, the online encyclopedia community, is written and maintained by volunteers through a decentralized community structure. It has 325 language communities and each community sets guidelines and rules regarding writing styles, moderation, and user behaviors independently. Meanwhile, within each language community, coalitions of editors with common interests or expertise in a particular domain have established their local community norms [47].

While such a decentralized community structure plays a significant role in fighting against vandalism, Wikipedians have also developed algorithmic systems to facilitate moderation work [35]. One widely used algorithmic system is *the RCFilters*. The RCFilters, short for Recent Changes Filters, has been deployed in 23 Wikipedia language communities as default on the RecentChanges page since August 2018[1]. The RecentChanges page is a special page that lists information about the most recent edits on Wikipedia, including editor usernames, article titles, and comments [2]. As shown in Figure 1, the RCFilters at this page enables content moderators to flag edits that *may have problems*, *likely have problems*, or *very likely have problems*. These three filters have different trade-offs between precision (*the proportion of flagged edits that are actually damaging*) and recall (*the proportion of actually damaging edits that are flagged*) to accommodate workflows of as many moderators as possible.

The RCFilters is powered by the ORES, an algorithmic scoring service that offers real-time damaging predictions for edits. The ORES prediction model is based on gradient boosted decision trees and trained with human-labelled Wikipedia edits [30]. Three thresholds are chosen for three filters and edits are flagged accordingly after comparing their ORES predictions with the thresholds. In this article, we measure the effect of the RCFilters on moderation decisions and the disparate treatment between editor groups at scale in communities of varying sizes and local editing activity.

4 RESEARCH HYPOTHESES

4.1 Effects of the RCFilters on Content Removal

Before presenting results associated with community characteristics, we first establish that the RCFilters has an effect on Wikipedia's moderation work. This test is fundamental to our following analysis because its results indicate whether the moderation tool RCFilters is used in Wikipedia. Given flagged edits are more likely to be spotted and then closely examined by moderators, we hypothesize that flagged edits have a higher reverting probability compared to non-flagged ones.

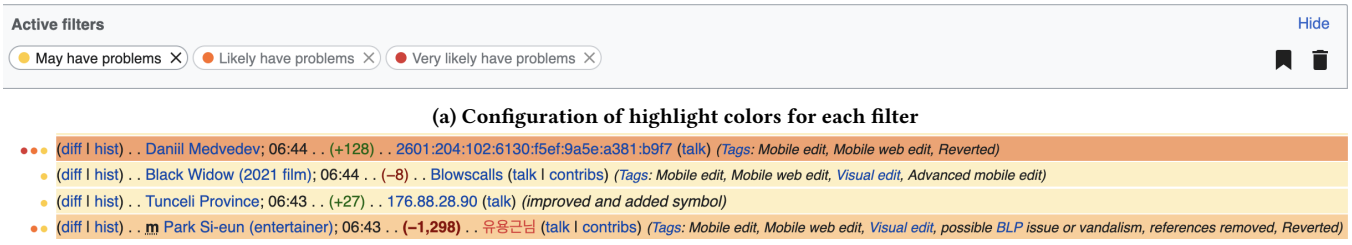
Analysis Check: *Edits that receive damaging scores immediately above flagging thresholds are more likely to be reverted compared to edits that receive damaging scores immediately below flagging thresholds.*

4.2 Influence of Community Sizes

Online communities like Reddit, Wikipedia, Discord, and Twitch have witnessed the proliferation of sub-communities built based on similar interests or languages. Although sub-communities share similar underlying technological infrastructures, they develop algorithms and moderate content independently [8, 9, 59]. Large communities are shown to rely more on technical approaches to moderation, including automated algorithms and machine-assisted systems. A field evaluation of Twitch moderation observed that

[1] RCFilters is part of the project Edit Review Improvements. See https://www.mediawiki.org/wiki/Edit_Review_Improvements/New_filters_for_edit_review

[2] This following link is directed at the RecentChanges page in English Wikipedia: https://en.wikipedia.org/wiki/Special:RecentChanges



(a) Configuration of highlight colors for each filter

(b) List of filtered and flagged edits

Figure 1: Screenshot of the RCFilters on the RecentChanges page in English Wikipedia. Different highlight colors inform moderators of the algorithmic damaging predictions. Note if an editor is anonymous, its corresponding username will be an IP address, as shown by the first line of the list.

there are four times more moderation actions taken by bots than humans in large channels in contrast to a roughly even number by bots and humans in medium and small channels [59].

There are several possible explanations for the heavy reliance of larger communities on moderation algorithms. First, sociologists suggest that individuals in large communities are less likely to form personal relationships and to coordinate with others efficiently [19]. Besides, large online communities are struggling to moderate huge volumes of everyday content [44]. Consequently, instead of focusing on removing all damaging content in a context-sensitive way, moderators in larger communities are more concerned about removing the most damaging content efficiently, in which moderation algorithms are highly competent.

Embedded in the Wikipedia front page, the tool RCFilters can efficiently identify and flag undesirable edits, and adapts well to the workflow of moderators in large Wikipedia communities. Therefore, we hypothesize that this flagging system will have a greater impact on the moderation work in larger Wikipedia communities.

Hypothesis 1: *The effect of the moderation tool (i.e., the difference in the reverting rate for edits above versus below flagging thresholds) will be greater in large communities than in small ones.*

4.3 Influence of Local Editing Activity

In addition to employing or appointing a small set of community members as moderators, online communities also increasingly adopt a decentralized moderation approach and rely on local community members to regulate content. In some cases, community members become part of a local community simply because they view the same post or video. For example, Slashdot users are enabled to down-vote unwelcome content in the comment section of an article, and viewers on the Twitter and YouTube platform can also report vandalism content. In Wikipedia, members vigilantly monitor recent edits to articles which they have contributed to before and remove damaging edits when needed [64]. On the other hand, local communities such as Facebook Groups and Wikipedia Projects have a clearer organization structure and establish local community norms. Specifically, Wikipedia Projects consist of editors with common interests or expertise in a particular domain and provide opportunities for editors to peer review each other's edits as well as protect their interesting articles from intruding zealotry and vandalism [20].

Prior research has underscored the indispensable role local communities play in regulating content. Lampe and Resnick argued that the distributed moderation system in Slashdot can quickly and consistently separate high and low quality comments [49]. Geiger underscored the collective and bottom-up modes of computationally assisted moderation through a field study of the bot-based blocklists in Twitter [22]. Kittur et al. showed joining a WikiProject encourages people to spend more efforts in reverting vandalism for articles within their affiliated WikiProject domain [47].

The proliferated local communities in Wikipedia mean that edits are always placed under close examination by local editors who are either moderators of a local WikiProject or individual moderators keeping an eye on articles out of their interest. Therefore, we hypothesize that the flagging system will have less influence on articles that have more local editing activity.

Hypothesis 2: *The effect of the moderation tool (i.e., the difference in reverting rates for edits above versus below thresholds) will be smaller in articles with more local editing activity compared to articles with less local editing activity.*

4.4 Impact of Moderation Algorithms on the Disparate Treatment towards Anonymous Editors

Online communities are always concerned about whether they should require would-be contributors to create accounts. Researchers have argued that there is no clear connection between anonymity and vandalism [65]. Although requiring an account will deter a large proportion of low-quality participation, the vast majority of rejected contributions are high-quality [37]. On the other hand, the creation of an account helps establish an identity in the community and enables editors to cooperate with other people more efficiently [7, 66]. Editing without a stable identifier is therefore considered as a bad form and more so if we notice that anonymity online will lead to misbehaviour like spam and hate speech [40]. In reality, anonymous edits are more closely scrutinized by moderators in Wikipedia [65].

Recently, Teblunthuis et al. investigated how the effect of moderation algorithms depends on the anonymous status of editors and found that, in the broadest strokes, algorithms can reduce discrimination against anonymous editors. However, they also acknowledged that, when it comes to details, algorithms have a much more

contingent and mixed influence [62]. In our analysis, we further explore how the effect of the algorithm upon the disparity between anonymous and registered editors depends on the community size and the local editing activity.

Without algorithmic flags, anonymous edits are more closely scrutinized [65] and edits by registered editors are more likely to go unnoticed. In contrast, the algorithm behind the RCFilters uses a set of language features to score the damaging likelihood of each edit, allowing the RCFilters to surface problematic edits by all kinds of editors. Therefore, we hypothesize that, while anonymous edits - regardless of being flagged or not - will continue to be closely scrutinized, edits by registered editors will only be more carefully reviewed by moderators after being flagged. This means that the effect of the moderation tool on anonymous editors will be smaller than that on registered editors. In other words, the content moderation tool can promote fairness between anonymous and registered editors.

Hypothesis 3: *The moderation tool can reduce the disparate treatment between anonymous and registered editors (i.e. the difference in the effect of the moderation tool for edits made by anonymous editors and that by registered editors).*

In communities of varying sizes and local editing activity, the relationship between algorithmic flags and the disparity between editor groups is more complicated. If we believe that the RCFilters can promote fairness between editor groups in general, whether the RCFilters is frequently used in a community will greatly affect whether the tool increases fairness between editor groups in that community. Since we hypothesize that the RCFilters will be relied more in communities of larger sizes and with less local editing activity in our Hypothesis 1 and 2, we hypothesize that the RCFilters will also help alleviate the disparate treatment between editor groups in these two kinds of communities.

Hypothesis 4a: *The effect of the moderation tool on the disparate treatment between anonymous and registered editors will be smaller in larger communities.*

Hypothesis 4b: *The effect of the moderation tool on the disparate treatment between anonymous and registered editors will be smaller in communities of less local editing activity.*

5 DATA AND MEASUREMENT

5.1 Data Preparation

We obtained our data from the publicly archived data of all Wikipedia communities[3]. In particular, we analyzed data from 23 Wikipedia language editions where the RCFilters was enabled by default in the RecentChanges page between 2019 and 2020. We focus on the *revision* in our analysis, which is defined as a single edit to an article by a participant, who can be either a human editor or an automated robot. We then apply the following exclusion criteria and present the descriptive statistics of our dataset in the Table 1.

- First, since we plan to compare the use of all three filters of the RCFilters, we exclude Wikipedia language editions where there are less than three filters, such as France and Romanian Wikipedia. We are then left with 17 Wikipedia language editions.
- Because we aim to explore how the flagging system affects the human decision-making process, we remove revisions by robots from our dataset.
- Since we plan to compare edits immediately below and above the threshold, we restrict our analysis to revisions whose ORES predictions are within 0.03 of the thresholds. We will explain our quasi-experimental design in the Analysis section.

5.2 Measurement

Measurement of Reverts Following a large body of research on Wikipedia, we adopt identity reverts as the measurement of reverts in our analysis [31, 33, 34, 37]. A revision is considered as being reverted if and only if a second user's revision restores the page to its state before the first revision. The Wikipedia interface enables moderators to perform identity reverts with a single click.

Measurement of Community Sizes Our first hypothesis *H1* concerns how the community size might influence the use of the flagging system. Since the number of new edits represents the workload of moderation, we measure the size of a Wikipedia community by the month-average number of new edits from 2019 to 2020 obtained from the Wikimedia Statistics website[4]. We further classify all communities into three categories: small, medium, and large wikis by comparing their sizes with 100,000 and 500,000 respectively. Among 17 Wikipedia language editions, there are 4 large wikis, 6 medium wikis, and 7 small wikis.

Measurement of Local Editing Activity For our second hypothesis *H2*, for each revision, we count the number of local active editors as the number of unique editors who had edited at least once to this revision's corresponding article during a month period (half a month before and after this revision was submitted). For each language edition, a revision is said to be "very active" if its number of active editors ranks the top third of all revisions, to be "moderately active" if that ranks between one-third and two-thirds, and to be "less active" otherwise.

Measurement of Anonymity Our third and fourth hypotheses require us to identify whether the editor is anonymous, which has already been available in the archived dataset.

6 ANALYSIS

Ideally, to identify the causal relationship between algorithmic flags and reverts, we should compare the changes in the reverting probability of flagged edits to non-flagged edits, and guarantee that the flagging treatment is randomly assigned to edits. However, the archival data of Wikipedia are observational. Besides, whether an edit is flagged is determined by RCFilters thresholds and ORES predictions calculated from a collection of language features. These language features, such as the number of bad or informal words, can influence both the treatment and the outcome, and therefore are potential confounding factors in our analysis.

To ameliorate the confounding problem, we adopt the Regression Discontinuity Design (RDD) to approximate randomization. As one analytic method in observational causal inference, the RDD seeks to

[3]https://dumps.wikimedia.org/

[4]There are other metrics available, for example, the week-average and the bi-week-average number of new edits. We find that these metrics are strongly correlated and therefore only choose one of them.

Filter	Group	N.Obs	R.Pr	A.Pr
May have problems	Small wikis	459239	0.02	0.05
	Medium wikis	315726	0.06	0.27
	Large wikis	1897603	0.12	0.45
Likely have problems	Small wikis	25229	0.36	0.86
	Medium wikis	143988	0.37	0.92
	Large wikis	814557	0.37	0.87
Very likely have problems	Small wikis	14976	0.74	0.95
	Medium wikis	90313	0.77	0.97
	Large wikis	303388	0.76	0.91

Filter	Group	N.Obs	R.Pr	A.Pr
May have problems	Less active	1018536	0.06	0.38
	Moder. active	901337	0.11	0.36
	Very active	752695	0.13	0.36
Likely have problems	Less active	402316	0.26	0.89
	Moder. active	303483	0.44	0.88
	Very active	277975	0.46	0.86
Very likely have problems	Less active	131623	0.65	0.94
	Moder. active	153256	0.83	0.93
	Very active	123798	0.81	0.91

Table 1: Descriptive Statistics of the Dataset: number of observations, proportion of reverted revisions, and proportion of anonymous editors in each subgroup and each threshold

estimate the causal effect by comparing outcomes above and below the threshold that represents an exogenous change in a "forcing variable" [50, 54]. Given the selection of a threshold for a continuous forcing variable is arbitrary, observations immediately below and above the threshold can be considered as the control group and the treated group respectively after adjusting the underlying trend of the forcing variable. For example, to identify the effect of passing an exam, we could compare students who barely passed and failed after adjusting the effect of the minor difference between their exam grades.

In virtue of RDDs, we measure the effect of the flagging system on the reverting probability by modeling the ORES prediction as the forcing variable and the set of thresholds arbitrarily chosen for the RCFilters as the threshold variable.

6.1 Model Variables

We first introduce variables we will use in our regression models. In the following, the subscript r denotes that the variable varies across each revision, the subscript w denotes that the variable varies across each Wikipedia language edition, and the subscript j denotes that the variable varies across each filter.

- $Revert_r$ This dichotomous variable indicates whether a revision is reverted or not.
- $Dist_{rj}$ For each revision, this continuous variable measures the distance between its ORES prediction and a filter's threshold. Since the threshold varies across Wikipedia language editions, $Dist_{rj}$ is calculated as follows: $Dist_{rj} = score_r - threshold_{jw}$
- $Flagged_{rj}$ This dichotomous variable indicates whether a revision is flagged or not by a filter. $Flagged_{rj} = 1[Dist_{rj} > 0]$
- $WikiSize_r$ This categorical variable indicates the size of the Wikipedia community the revision belongs to. It can be one of the following three values: *small*, *medium*, *large*.
- $ActiveEditor_r$ This categorical variable indicates the number of active editors by whom a revision will be monitored. It can take the following three values: $less\ active$, $moderately\ active$, $very\ active$.
- $EditorRole_r$ This categorical variable indicates whether a revision is submitted by an anonymous editor. It can take the following two values: *anonymous*, $registered$.

6.2 Regression Models

Our analysis check examines the overall influence of algorithmic flags on the decision-making process, which we estimate by the Equation 1. For each filter, the collection of middle two terms $(\alpha_j Dist_{rj} + \beta_j Dist_{rj} \times Flagged_{rj})$ represents how the ORES prediction itself predicts the reverting probability. Following established approaches to RDDs, we allow for a change in slope at the cutoff [50, 54]. We also include the term c_{wj} to account for the differences of the average reverting probability across different Wikipedia communities and filters. Finally, the coefficient τ_j denotes the effect of the flagging system on the reverting probability, which is the focal point in our analysis.

$$log\left(\frac{Revert_r}{1 - Revert_r}\right) = \tau_j Flagged_{rj} + \alpha_j Dist_{rj} + \beta_j Dist_{rj} \times Flagged_{rj} + c_{wj} \quad (1)$$

On the basis of the Equation 1, we further include interaction terms to examine the hypotheses as listed below. For convenience of exposition, we refer to the right-hand side of the Equation 1 as the *main variables*.

- For Hypothesis *1*, we have Model 1: examine the interaction between the main variables and $WikiSize_r$.
- For Hypothesis *2*, we have Model 2: examine the interaction between the main variables and $ActiveEditor_r$.
- For Hypothesis *3*, we have Model 3: examine the interaction between the main variables and $EditorRole_r$.
- For Hypothesis *4a*, we have Model 4a: further allow the two-way interaction terms in *Model 1* to vary across the $EditorRole_r$.
- For Hypothesis *4b*, we have Model 4b: further allow the two-way interaction terms in *Model 2* to vary across the $EditorRole_r$.

Table 2 and Table 3 then present the analysis results for each model.

Analysis Check: Overall Effect		
Filter	Est.	Std.Err
May have problems	0.02**	0.01
Likely have problems	0.03**	0.01
Very likely have problems	0.18*	0.02

Model 1: Community Size			
Filter	Group	Est.	Std.Err
May have problem	Small wikis	1.26***	0.05
	Medium wikis	-0.06*	0.08
	Large wikis	-0.02*	0.01
Likely have problem	Small wikis	-0.01	0.05
	Medium wikis	0.03	0.03
	Large wikis	0.04	0.01
Very likely have problem	Small wikis	0.08	0.08
	Medium wikis	0.08*	0.02
	Large wikis	0.23***	0.02

Model 2: Local Editing Activity			
Filter	Group	Est.	Std.Err
May have problems	Less active	0.08***	0.02
	Moder. active	0.01	0.01
	Very active	0.01	0.01
Likely have problems	Less active	0.06***	0.01
	Moder. active	0.04*	0.01
	Very active	0.01	0.02
Very likely have problems	Less active	0.19***	0.03
	Moder. active	0.18***	0.03
	Very active	0.15***	0.03

Model 3: Anonymity			
Filter	Group	Est.	Std.Err
May have problems	Anonymous	-0.02	0.01
	Registered	0.07***	0.01
Likely have problems	Anonymous	0.04***	0.01
	Registered	0.03	0.02
Very likely have problems	Anonymous	0.18***	0.02
	Registered	0.16*	0.03

Table 2: Regression results. Regression predicting the effects of the flagging system on the reverting probability for each filter and each subgroup, i.e., the coefficient τ_j in the regression model (*p<0.05; **p<0.01; ***p<0.001)

7 FINDINGS

7.1 Effects of RCFilters on Content Removal

Before presenting results about the community size and local editing activity, we examine the overall effect of the RCFilters on revisions. In support of our analysis check and prior research, the Analysis Check Model in Table 2 predicts that being flagged by any filter significantly increases the reverting probability of a revision. Figure 2 illustrates that the "very likely have problems" filter contributes to an approximately 3.0% change in the reverting probability.

7.2 Influence of Community Sizes

For the "may have problems" filter, Model 1 in Table 2 predicts that being flagged significantly increases the reverting odds of a revision by 1.26 (95% CI: [1.17, 1.35]) in small Wikipedia communities. By contrast, the effects in medium and large communities are very close to zero (Est.=-0.06, 95% CI: [-0.12, 0.00]; Est.=-0.02, 95% CI: [-0.04, 0.00] respectively). This disparity is illustrated by Figure 3a from which we can observe an approximately 2.4% jump of the reverting probability for small communities but a slight drop for medium and large ones.

However, at the "very likely have problems" filter, we observe a significant flagging effect in medium wikis (Est.=0.08, p<0.05, 95% CI: [0.02, 0.14]) and large wikis (Est.=0.23, p<0.001, 95% CI: [0.19, 0.27]) but no significant effect in small wikis (Est.=0.08, 95% CI: [-0.08, 0.24]). Figure 3b shows that the change in the reverting probability in large wikis is around 3.6%, more significant than the changes in medium and small wikis. At the "likely have problems" filter, we find that the flagging system in large wikis contributes to an increase of the reverting logit between 2% and 5%, but we have less confidence in the flagging effect in medium wikis (Est.=0.04, 95% CI: [-0.01, 0.08]). We do not notice a significant effect in small wikis.

These results demonstrate that the impact of the flagging system varies across communities of different sizes. While flagging an edit as "may have problems" significantly increases its probability of revert in small communities, the flagging effect is limited in medium and large communities. Conversely, for the "very likely have problems" filter, flagged edits are significantly more likely to be reverted than non-flagged edits in medium and large communities, despite no detectable effect in small communities. Therefore, the **Hypothesis 1** is only supported for the "likely have problems" and "very likely have problems" filters but not for the "may have problems" filter.

7.3 Effects of Local Editing Activity

Model 2 in Table 2 tests the hypothesis *2* that whether local editing activity will mediate the effect of the RCFilters on moderation decisions. We observe that the flagging effects of the "may have problems" and "likely have problems" filters are less significant on revisions made on moderately active or very active articles. As visualized in Figure 4a, being flagged by the "may have problems" filter increases the reverting probability by around 0.5% for revisions of the "less active" group. However, the effect is less detectable for articles of the other two groups.

Surprisingly, for the "very likely have problems" filter, we do not observe significant differences in the flagging effect on revisions of different local editing activity groups. The "very likely have problems" filter increases the reverting odds for revisions of any

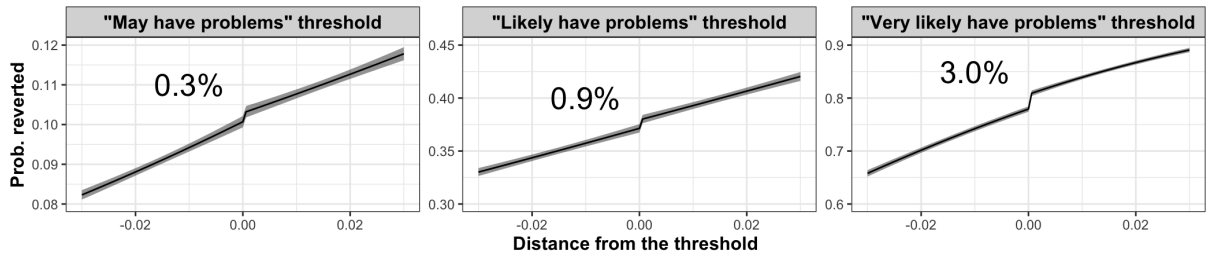
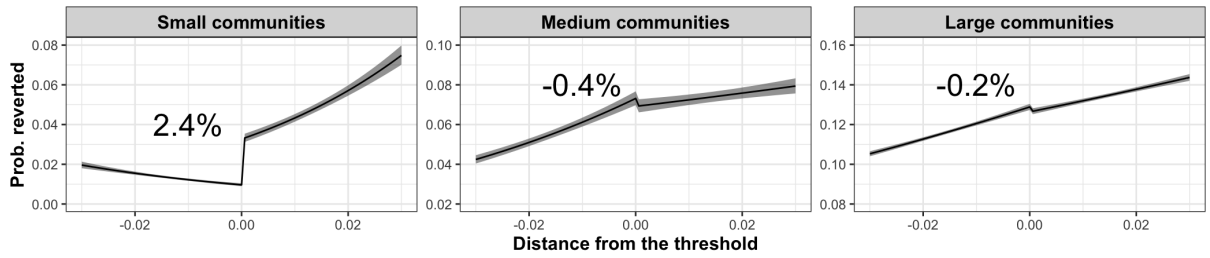
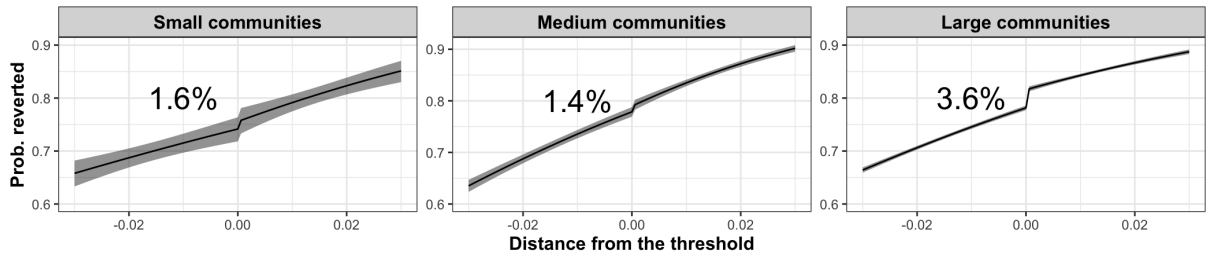


Figure 2: Effects of the RCFilters on content removal for each threshold. The x-axis is the distance away from the threshold and y-axis is the reverting probability with the 95% credible interval. We also estimate the jump at the cutoff.



(a) "May have problems" threshold



(b) "Very likely have problems" threshold

Figure 3: Effects of the RCFilters on content removal for each community size and threshold. The x-axis is the distance away from the threshold and the y-axis is the reverting probability with its 95% credible interval. We also estimate the jump at the cutoff.

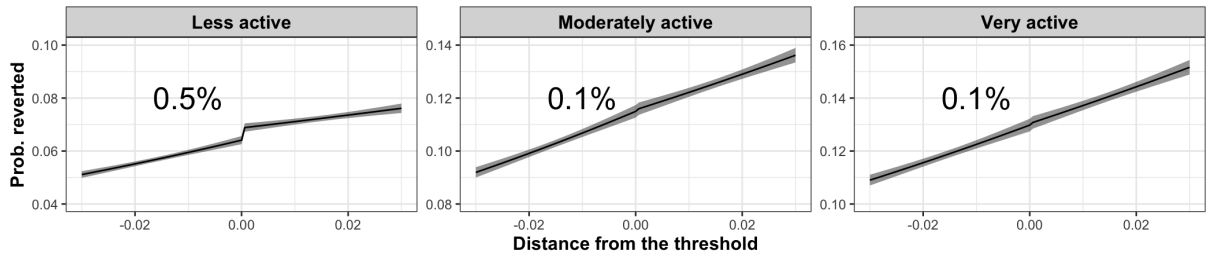
local editing activity group. Figure 4b shows that the change in the reverting probability is around 4.1% for the "less active" group, 2.3% for the "moderately active" group, and 2.3% for the "very active" group.

To summarise, for "may have problems" and "likely have problems" filters, the flagging effects are stronger for less active articles. Conversely, being flagged as "very likely have problems" significantly increases the reverting odds of revisions in all groups of articles. Therefore, our **Hypothesis 2** only holds for the first two filters.

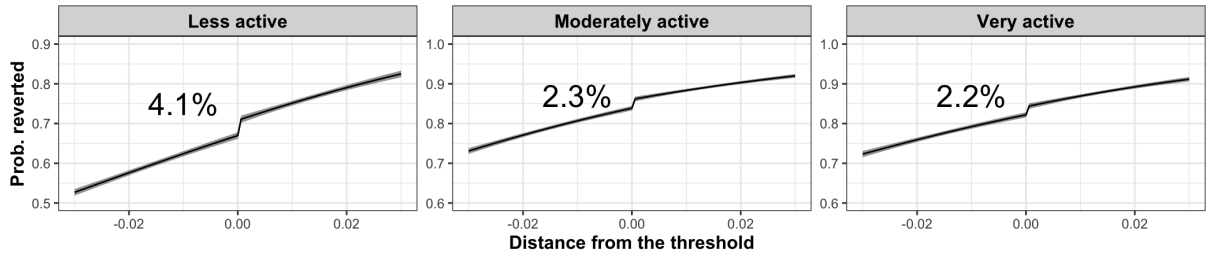
7.4 Effects of Anonymity

Model 3 in Table 2 shows the flagging effect on the disparate treatment between anonymous and registered editors. Consistent with [62], we find that the flagging effect of the "may have problems" filter for anonymous editors is significantly lower than that for registered editors (Est.Diff=-0.09, 95% CI: [-0.12, -0.05]). Our findings differ from [62] in that we do not detect significant differences between these two editor groups for the "likely have problems" and "very likely have problems" filters. Our **Hypothesis 3** is only supported for the "may have problems" filter. Therefore, we focus on the "may have problems" filter in the following and present the results for the other two filters in the Appendix.

7.4.1 Community Sizes. Model 4a in Table 3 shows that the flagging system has a divergent influence on the disparate treatment between anonymous and registered editors in communities of different sizes. In small communities, we find that the flagging effect on anonymous editors is significantly smaller than that on registered editors (Est.Diff=-0.69, 95% CI: [-0.87, -0.50]). On the contrary, the difference between two editor groups in medium communities



(a) "May have problems" threshold



(b) "Very likely have problems" threshold

Figure 4: Effects of the RCFilters on content removal for each level of local editing activity and threshold. The x-axis is the distance away from the threshold and the y-axis is the reverting probability with the 95% credible interval. We also estimate the jump at the cutoff.

is opposite in sign (Est.Diff=0.17, 95% CI: [0.05, 0.29]). For large communities, we do not detect significant differences between these two groups (Est.Diff=-0.02, 95% CI: [-0.06, 0.01]).

These findings suggest a dynamic relationship between algorithmic flags and the disparity between anonymous and registered editors: the flagging system reduces the disparity in small communities, increases it in medium communities, and has no detectable effect on it in large communities. Therefore, our **Hypothesis 4a** is not supported.

7.4.2 Local Editing Activity. By contrast, Model 4b in Table 3 indicates a more consistent impact of algorithmic flags on the disparity between editor groups. For edits in the "less active" and the "very active" group, the flagging effect is significantly weaker for anonymous editors than for registered editors (Est.Diff=-0.15, 95% CI: [-0.22, -0.09]; Est.Diff=-0.10, 95% CI: [-0.15, -0.04] respectively). For revisions in the "moderately active" group, we have less confidence in the difference between editor groups because the 95% credible interval includes zero (Est.Diff=-0.04, 95% CI: [-0.10, 0.02]). Our results indicate that the flagging system alleviates the bias against anonymous editors for less active or very active articles but has no detectable effect on moderately active articles. Our **Hypothesis 4b** is not supported. We attempt to explain the divergence between our hypotheses and findings in the Discussion.

8 DISCUSSION

8.1 Community Characteristics Can Shape the Use of Moderation Algorithms

A large body of research seeks to understand how groups integrate new technologies in various organizational setting. Orlikowski's concept of technological frames suggests that group members' shared "assumptions, expectations, and knowledge to understand technology" profoundly structure the way new technology is adopted and used [55]. Through the lens of this concept we discuss how community characteristics shape the use of moderation algorithms in Wikipedia.

Community Sizes We find that moderators in larger communities tend to use the "very likely have problems" filter and possibly the "likely have problems" filter, while moderators in small communities tend to use the "may have problems" filter. Given these filters only differ in their trade-offs between recall and precision, our findings indicate that moderators in large communities prioritise removing extremely damaging edits as quickly as possible, while moderators in small communities prioritise removing all edits that are harmful to varying degrees. A possible explanation is that vandalism in large communities is more widespread and detrimental due to their extensive publicity. Moderators therefore focus on the efficiency of moderation tools and adopt them primarily to facilitate vandalism fighting.

Local Editing Activity Our findings reveal that the interaction between local editing activity and the flagging system varies across filters. Being flagged as "may have problems" only impacts revisions on less active articles, while being flagged as "very likely have problems" increases the reverting odds by at least 1% regardless of the number of active editors. These findings suggest moderators in local communities use the moderation tools differently. Owing to the limited number of new edits in local communities and the closer relationship between community members, local moderators feel less stressful about vandalism and can coordinate more efficiently.

Model 4a: Anonymity × Community Size			
Group	Editor role	Est.	Std.Err
Small wikis	Anonymous	0.26***	0.08
	Registered	0.95***	0.06
Medium wikis	Anonymous	0.06	0.04
	Registered	-0.11**	0.04
Large wikis	Anonymous	-0.03*	0.01
	Registered	-0.01	0.01

Model 4b: Anonymity × Local Editing Activity			
Group	Editor role	Est.	Std.Err
Less active	Anonymous	0.01	0.02
	Registered	0.16***	0.03
Moderately active	Anonymous	-0.02	0.02
	Registered	0.02	0.02
Very active	Anonymous	-0.04	0.02
	Registered	0.05**	0.02

Table 3: Regression predicting the effect of the "may have problems" filter on the reverting probability for each editor role and each subgroup, i.e., the coefficient τ_j in the regression model (*p<0.05; **p<0.01; ***p<0.001)

Consequently, they rely less on machine-assisted systems and review edits more carefully. On the other hand, local moderators still find the "very likely have problems" filter useful in that it helps them identify and remove extremely harmful edits more quickly. They can then review less harmful edits more carefully in terms of their quality and alignment with local community norms.

8.2 The Effect of Moderation Algorithms on Fairness Varies across Communities

Prior research has offered disparate predictions about the impact of algorithms on fairness [5, 6, 17, 28, 39, 42]. Here we argue that the relationship between moderation algorithms and fairness is more complicated. In Wikipedia, whether and to what extent the algorithm promotes fairness in the community depends on a variety of practical factors, especially the moderation trade-offs of the algorithm and community characteristics.

Moderation Trade-offs Our findings are different from those of Teblunthuis et al. in that we demonstrate that the overall effect of the RCFilters on the disparate treatment between editor groups is only significant for the "may have problems" filter. One possible reason is that the other two filters of higher thresholds choose higher accuracy in the trade-off between accuracy and recall. Since very damaging edits are largely made by anonymous editors, in order to optimize accuracy, these two filters primarily flag anonymous edits instead of surfacing damaging edits by all kinds of editors. As we can see from the Table 1, among edits flagged by the "very likely have problems" filter, anonymous edits account for 95%. Hence, these two filters have little impact on the fairness between editor groups.

Community Characteristics We find that the "may have problems" filter reduces the disparity in small communities, increases it in medium communities, and has no detectable influence on it in large communities. We also find that the "may have problems" filter alleviates discrimination against anonymous editors regardless of the number of active editors in the local community, and the effect in the less active communities is the most significant. Since we have demonstrated before that this filter is *most* frequently used only in small and less active communities, its significant effect on fairness in these two communities indicates the ability of the moderation algorithm to promote fairness. Further, even this filter is less frequently used in moderately or very active communities, it is still able to promote fairness between editor groups. Surprisingly, our findings suggest that the "may have problems" filter reinforces the bias against anonymous editors in medium Wikipedia communities. Since the ORES system has selected a different set of language features and used a different training dataset in each language edition, one possible reason is that the algorithm performances on anonymous and registered edits differ in communities of varying sizes.

8.3 Design Implications for Future Human-centered Moderation Algorithms

Our findings have important design implications for future designers of social media applications and online platforms. **Empowering moderators to adapt moderation tools to their workflows**. Our findings echo advocates of prior research that stakeholders should have the authority to determine trade-offs between critical criteria of an algorithmic system so that they can fully take advantage of the system to facilitate their moderation work [61]. In practice, we suggest future moderation systems present the system performances (e.g. accuracy, false-positive rate etc.) to moderators and empower them to choose the most desirable threshold according to their workflows and community characteristics.

Supporting moderation work of communities of different sizes. Our findings highlight the community size as an important factor that determines the priority of moderation tasks. While most moderation algorithmic systems are designed to help scale manual moderation [4, 27] and therefore better serve the priorities of large communities, system developers should investigate the special needs of small online communities and support them accordingly. For example, moderators in small communities collaborate with each other more often, which underscores the needs to record the moderation history of each person and support more efficient task allocation [19].

Besides, while we find that small communities tend to use the "may have problems filter", we also notice some room for improvements of the algorithmic performance of this filter in small communities. As shown in the figure 3a, for edits with ORES scores below the "may have problems" threshold, a higher ORES prediction score predicts a lower reverting probability in small communities,

contrary to a higher reverting probability in medium and large communities. This suggests that the ORES algorithm performs better in medium and large communities than in small communities.

Motivating and empowering local community members to moderate content. We present another side of the moderation work in online communities: editors of local communities build another layer of the immune system in addition to fully autonomous robots and centralized moderators [35]. Local editors can review content in a context-sensitive manner by carrying local community norms in mind. Online platforms should consider how to design their platform structures to incentivize local editors to participate in the fights against vandalism. One possible approach is to reinforce people's identity as a member of a local community. Kittur et al. showed that joining a WikiProject encourages people to spend more efforts in reverting vandalism for articles within their affiliated WikiProject domain [47]. Therefore, we suggest system developers recommend relevant interest groups to local editors or offer dedicated local editors a badge or an award to give them a sense of community.

9 LIMITATION AND FUTURE WORK

Our findings are subject to several threats to validity associated with the regression discontinuity design we adopt to draw causal inferences. Formally, in RDD models, if optimizing agents do not have *precise* control over the forcing variable, then the variation in the treatment will be as good as randomized in a neighborhood around the discontinuity threshold [50, 54]. In an ideal RDD setting, individuals are allowed to exert some control over the forcing variable but they are assumed to be incapable of self-selection into treatment or control groups.

Although this assumption of the RDD method is fundamentally unverifiable, we believe that the RCFilters is an ideal setting to apply the RDD method for the following reasons. One significant challenge to the assumption is the hypothetical possibility of potential vandals' handcrafting their damaging edits to avoid algorithmic flags. The ORES system provides detailed information about a prediction through its public API and allows users to inject their features and observe how the prediction changes accordingly. However, Wikipedia's complex moderation network means that a successful evasion of the RCFilters flags would not guarantee a successful vandalism act. The great efforts required and the low payoff of such actions suggest they are unrealistic and very unlikely to be widespread [62].

Secondly, to test the assumption of RDDs, we examine whether the treatment and control groups are similar in their observed baseline characteristics that are determined prior to the treatment [50]. We expect there is no abrupt jump or drop of predetermined characteristics across the cutoff. We select the following baseline features shared by revisions in all Wikipedia language editions to test our assumption: the number of external links of each revision and its parent article, the length of each revision and its parent article, and the length of the longest repeated characters. We substitute the left-hand side of the Equation 1 with each feature and present the regression results in the Appendix. In support of our assumption, we do not find detectable differences of these predetermined characteristics across the threshold.

Besides, since we aimed to quantify the effect of the RCFilters on moderation decision-making in community of different characteristics, we only focused on moderation decisions between 2019 and 2020 and did not investigate how this effect changes over time. It is possible that, as community members continue to use the RCFilters, their perception of the moderation system and their interaction with this system changes, which resulted in varying effects of the RCFilters at different time periods. Future work could more carefully investigate how machine-assisted moderation decisions might vary across time.

Finally, in order to provide a detailed case study of how machine-assisted moderation systems are actually used, we narrowed our analysis to the Wikipedia context, and therefore our findings are inevitably closely related to the setting of the RCFilters and the Wikipedia community. However, we believe our findings are still able to provide important insights for understanding content moderation practice in a broader context for the following reasons. Firstly, the proliferation of channels and sub-communities in various online platforms like Reddit, Twitch, and Discord highlights the need to understand the variance of moderation needs and practice among these sub-communities. While prior research has revealed the greater moderation challenges of large communities [44] and their heavier reliance on automated moderation approach [43, 59], we highlighted the variance of moderation goals across communities of different sizes and the importance of designing moderation systems accordingly. Secondly, numerous local communities have grown on online platforms and started to moderate their content in a community-based approach [47, 49]. We revealed the critical role local communities play in protecting their community against vandalism and the need to design machine-assisted moderation systems for local communities.

10 CONCLUSION

In this research, we examine several factors that shape the use of a machine-assisted moderation system named RCFilters in the sociotechnical context of Wikipedia through a regression discontinuity design. We found that 1) larger communities tend to use predictive systems to identify the most damaging edits, while smaller communities tend to use them to identify any edit that could be damaging; 2) predictive systems are used less in content areas where there are more local editing activities; 3) predictive systems have mixed effects on reducing disparate treatment between anonymous and registered editors in different communities.

ACKNOWLEDGMENTS

We would like to thank Zhiwei Steven Wu, Nathan TeBlunthuis, Aaron Halfaker, members of the HCII at Carnegie Mellon University, and our anonymous reviewers for their help in this project. This work was supported by the National Science Foundation (NSF) under Award No.2001851, 2000782, 1952085, and the NSF Program on Fairness in AI in collaboration with Amazon under Award No.1939606.

REFERENCES

[1] Adrian Chen. 2014. The Laborers Who Keep Dick Pics And Beheadings Out Of Your Facebook Feed. "https://www.wired.com/2014/10/content-moderation/". [Online; accessed 19-August-2021].

[2] Andrew Liptak. 2017. IMDb is closing its message boards. "https://web.archive.org/web/20190410195135/https://www.theverge.com/2017/2/3/14501390/imdb-closing-user-forums-comments". [Online; accessed 07-Sep-2021].

[3] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[4] Monika Bickert. 2018. Publishing our internal enforcement guidelines and expanding our appeals process. https://about.fb.com/news/2018/04/comprehensive-community-standards/

[5] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*. Springer, 405–415.

[6] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[7] Susan L. Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work* (2005), 1–10. https://doi.org/10.1145/1099203.1099205

[8] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–30. https://doi.org/10.1145/3359276

[9] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

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

[11] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[13] Bo Cowgill. 2019. Bias and productivity in humans and machines. *Columbia Business School Research Paper Forthcoming* (2019).

[14] Paul B de Laat. 2016. Profiling vandalism in Wikipedia: A Schauerian approach to justification. *Ethics and Information Technology* 18, 2 (2016), 131–148.

[15] Paul B. de Laat. 2016. Profiling vandalism in Wikipedia: A Schauerian approach to justification. *Ethics and Information Technology* 18, 2 (2016), 131–148. https://doi.org/10.1007/s10676-016-9399-8

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[17] Michael D Ekstrand and Daniel Kluver. 2021. Exploring author gender in book rating and recommendation. *User modeling and user-adapted interaction* (2021), 1–44.

[18] Elizabeth Jensen. 2016. NPR Website To Get Rid Of Comments. "https://web.archive.org/web/20190723073441/https://www.npr.org/sections/publiceditor/2016/08/17/489516952/npr-website-to-get-rid-of-comments". [Online; accessed 07-Sep-2021].

[19] Scott L Feld. 1982. Social structural determinants of similarity among associates. *American sociological review* (1982), 797–801.

[20] Andrea Forte, Niki Kittur, Vanessa Larco, Haiyi Zhu, Amy Bruckman, and Robert E Kraut. 2012. Coordination and beyond: social functions of groups in open content production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 417–426.

[21] R Stuart Geiger. 2014. Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society* 17, 3 (2014), 342–356.

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

[23] R. Stuart Geiger and Aaron Halfaker. 2013. When the levee breaks: Without bots, what happens to wikipedia's quality control processes? *Proceedings of the 9th International Symposium on Open Collaboration, WikiSym + OpenSym 2013* (2013), 0–5.

[24] R. Stuart Geiger and David Ribes. 2010. The work of sustaining order in wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*. ACM Press, New York, New York, USA, 117. https://doi.org/10.1145/1718918.1718941

[25] R Stuart Geiger and David Ribes. 2010. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 117–126.

[26] R Stuart Geiger and David Ribes. 2011. Trace ethnography: Following coordination through documentary practices. In *2011 44th Hawaii international conference on system sciences*. IEEE, 1–10.

[27] Google. 2018. YouTube Community Guidelines enforcement in Google's Transparency Report for 2018. https://transparencyreport.google.com/youtube-policy/removals/

[28] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[29] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

[30] Aaron Halfaker and R. Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020). https://doi.org/10.1145/3415219 arXiv:1909.05189

[31] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System. *American Behavioral Scientist* 57, 5 (may 2013), 664–688. https://doi.org/10.1177/0002764212469365

[32] Aaron Halfaker, R. Stuart Geiger, and Loren Terveen. 2014. Snuggle: Designing for efficient socialization and ideological critique. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 311–320. https://doi.org/10.1145/2556288.2557313

[33] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A jury of your peers: Quality, experience and ownership in Wikipedia. *Proceedings of the 5th International Symposium on Wikis and Open Collaboration, WiKiSym 2009* (2009). https://doi.org/10.1145/1641309.1641332

[34] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. *WikiSym 2011 Conference Proceedings - 7th Annual International Symposium on Wikis and Open Collaboration* (2011), 163–172. https://doi.org/10.1145/2038558.2038585

[35] Aaron Halfaker and John Riedl. 2012. Bots and cyborgs: Wikipedia's immune system. *Computer* 45, 3 (2012), 79–82. https://doi.org/10.1109/MC.2012.82

[36] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[37] Benjamin Mako Hill and Aaron Shaw. 2020. The Hidden Costs of Requiring Accounts: Quasi-Experimental Evidence From Peer Production. *Communication Research* (2020), 0093650220910345.

[38] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 5 (2019), 1–35.

[39] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[40] Adam Joinson. 1998. Causes and implications of disinhibited behavior on the Internet. (1998).

[41] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139* (2016).

[42] Julia Angwin, Hannes Grassegger. 2017. Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. "https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms". [Online; accessed 06-Jan-2022].

[43] Charles Kiene and Benjamin Mako Hill. 2020. Who uses bots? a statistical analysis of bot usage in moderation teams. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–8. https://doi.org/10.1145/3334480.3382960

[44] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological frames and user innovation: exploring technological change in community moderation teams. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[45] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. Building Successful Online Communities: Evidence-Based Social Design.

[46] Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information Communication and Society* 20, 1 (2017), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

[47] Aniket Kittur, Bryan Pendleton, and Robert E Kraut. 2009. Herding the cats: the influence of groups in coordinating peer production. In *Proceedings of the 5th international Symposium on Wikis and Open Collaboration*. 1–9.

[48] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[49] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[50] David S Lee and Thomas Lemieux. 2010. Regression discontinuity designs in economics. *Journal of economic literature* 48, 2 (2010), 281–355.
[51] Maeve Duggan. 2017. Online Harassment 2017. "https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/". [Online; accessed 07-Sep-2021].
[52] Ariadna Matamoros-Fernández. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society* 20, 6 (2017), 930–946.
[53] Kathleen L Mosier and Linda J Skitka. 2018. Human decision makers and automated decision aids: Made for each other? In *Automation and human performance: Theory and applications*. CRC Press, 201–220.
[54] Richard J Murnane and John B Willett. 2010. *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.
[55] Wanda J Orlikowski and Debra C Gash. 1994. Technological frames: making sense of information technology in organizations. *ACM Transactions on Information Systems (TOIS)* 12, 2 (1994), 174–207.
[56] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993* (2017).
[57] Robert Peck. 2019. The Punishing Ecstasy of Being a Reddit Moderator. https://www.wired.com/story/the-punishing-ecstasy-of-being-a-reddit-moderator/
[58] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[59] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.
[60] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Vol. 1. ACM, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598
[61] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–14. https://doi.org/10.1145/3313831.3376783
[62] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2021. Effects of Algorithmic Flagging on Fairness. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27. https://doi.org/10.1145/3449130
[63] Tiziana Terranova. 2000. Free labor: Producing culture for the digital economy. *Social text* 18, 2 (2000), 33–58.
[64] Jennifer Thom-Santelli, Dan R Cosley, and Geri Gay. 2009. What's mine is mine: territoriality in collaborative authoring. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1481–1484.
[65] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, Vol. 6. ACM Press, New York, New York, USA, 575–582. https://doi.org/10.1145/985692.985765
[66] Fernanda B Viégas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. Talk before you type: Coordination in Wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE, 78–78.
[67] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[68] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[69] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–23. https://doi.org/10.1145/3274463

A APPENDIX

Model 4a: Anonymity × Community Size			
Group	Editor role	Est.	Std.Err
Small wikis	Anonymous	-0.02	0.06
	Registered	0.02	0.14
Medium wikis	Anonymous	0.04	0.02
	Registered	-0.08	0.08
Large wikis	Anonymous	0.04***	0.01
	Registered	0.04	0.03

Model 4b: Anonymity × Local Editing Activity			
Group	Editor role	Est.	Std.Err
Less active	Anonymous	0.06***	0.02
	Registered	0.03	0.04
Moderately active	Anonymous	0.03*	0.02
	Registered	0.06	0.04
Very active	Anonymous	0.02	0.02
	Registered	-0.02	0.04

Table 4: Regression predicting the effects of the "likely have problems" filter on the reverting probability for each editor role and each subgroup, i.e., the coefficient τ_j in the regression model (*p<0.05; **p<0.01; ***p<0.001)

Model 4a: Anonymity × Community Size			
Group	Editor role	Est.	Std.Err
Small wikis	Anonymous	0.08	0.08
	Registered	0.13	0.36
Medium wikis	Anonymous	0.09**	0.03
	Registered	-0.45**	0.17
Large wikis	Anonymous	0.22***	0.02
	Registered	0.27***	0.08

Model 4b: Anonymity × Local Editing Activity			
Group	Editor role	Est.	Std.Err
Less active	Anonymous	0.19***	0.03
	Registered	0.20	0.11
Moderately active	Anonymous	0.17***	0.03
	Registered	0.39**	0.13
Very active	Anonymous	0.17***	0.03
	Registered	-0.07	0.12

Table 5: Regression predicting the effects of the "very likely have problems" filter on the reverting probability for each editor role and each subgroup, i.e., the coefficient τ_j in the regression model (*p<0.05; **p<0.01; ***p<0.001)

The Number of External Links of a Revision		
Filter	Est.	Std.Err
May have problems	0.19	0.16
Likely have problems	0.11	0.26
Very likely have problems	-0.31	0.41
The Number of External Links of the Parent Article		
Filter	Est.	Std.Err
May have problems	-0.20	0.16
Likely have problems	0.12	0.26
Very likely have problems	-0.22	0.41
The Length of a Revision		
Filter	Est.	Std.Err
May have problems	-11.84	10.91
Likely have problems	-19.45	17.49
Very likely have problems	-45.97	27.68
The Length of the Parent Article		
Filter	Est.	Std.Err
May have problems	-12.16	10.91
Likely have problems	-18.18	17.48
Very likely have problems	-34.01	27.67
The Length of the Longest Repeated Char		
Filter	Est.	Std.Err
May have problems	0.07	0.24
Likely have problems	-0.08	0.39
Very likely have problems	0.75	0.62

Note: *p<0.05; **p<0.01; ***p<0.001

Table 6: Regression predicting the change of the predetermined characteristic at the cutoff for each threshold