Effects of Algorithmic Flagging on Fairness: Quasi-experimental Evidence from Wikipedia

NATHAN TEBLUNTHUIS*, University of Washington, USA
BENJAMIN MAKO HILL, University of Washington, USA
AARON HALFAKER†, Microsoft, USA

Online community moderators often rely on social signals such as whether or not a user has an account or a profile page as clues that users may cause problems. Reliance on these clues can lead to "overprofiling" bias when moderators focus on these signals but overlook the misbehavior of others. We propose that algorithmic flagging systems deployed to improve the efficiency of moderation work can also make moderation actions more fair to these users by reducing reliance on social signals and making norm violations by everyone else more visible. We analyze moderator behavior in Wikipedia as mediated by RCFilters, a system which displays social signals and algorithmic flags, and estimate the causal effect of being flagged on moderator actions. We show that algorithmically flagged edits are reverted more often, especially those by established editors with positive social signals, and that flagging decreases the likelihood that moderation actions will be undone. Our results suggest that algorithmic flagging systems can lead to increased fairness in some contexts but that the relationship is complex and contingent.

CCS Concepts: • **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**; **Social content sharing**; **Computer supported cooperative work**.

Additional Key Words and Phrases: sociotechnical systems; moderation; AI; machine learning; causal inference; peer production; Wikipedia; online communities; community norms; fairness;

ACM Reference Format:
Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2021. Effects of Algorithmic Flagging on Fairness: Quasi-experimental Evidence from Wikipedia. In *Proceedings of CSCW '21: Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '21).* ACM, New York, NY, USA, 27 pages.

1 INTRODUCTION

Online community moderators are responsible for reviewing the torrents of user-generated content for spam, vandalism, attacks, and other violations of community norms and rules. In many large online communities, a small number of moderators—often volunteers—will be responsible for reviewing thousands or millions of actions and taking steps to stop and mitigate problematic behaviors [30]. To help focus their attention within this deluge, moderators typically rely on social signals [21] that indicate that a user's contributions are made in good faith and of high quality [57]. Common signals include visible reputation scores, user profiles, experience, and registration status [10, 57]. For example, since new users are often more likely to engage in bad behaviors, moderators might scrutinize contributions from newcomers more closely [57, 80]. However, directing limited

*Part of this author's contributions to this paper were made while he was affiliated with the Wikimedia Foundation.
†The majority of this author's contributions to this paper was completed when he was affiliated with the Wikimedia Foundation.

CSCW '21, November 03–07, 2021, Toronto, Canada

moderation attention based on social signals can introduce unfairness through "overprofiling" that occurs when moderators focus their attention on users with signals associated with bad behaviors while ignoring others engaged in similar or worse behaviors [19]. For this reason, and because relying on social signals can still place enormous demands on limited moderator resources, online communities are increasingly adopting algorithmic flagging systems to direct moderators toward problematic actions [14, 33].

Although the consequences are very different, these systems share salient commonalities with algorithmic flagging systems used in employment, college admissions, and criminal justice. All of these systems use predictions of whether an outcome will occur to flag certain individuals as more or less likely sources of problems. All leave final decisions to a human judge. The use of these systems when people's lives are at stake has rightfully attracted criticism based of how algorithms engage in misrepresentation and discrimination [4, 12, 76]. On the other hand, advocates of algorithmic prediction in criminal justice argue that algorithms—even those that are measurably biased in their predictions—might still be less discriminatory than decisions made by biased human judges alone [55, 90].

Can algorithmic flagging systems in online community moderation similarly reduce reliance on social signals and lead to more fair outcomes? We aim to answer this question through a field evaluation of an algorithmic flagging system called RCFilters, which was deployed on 23 different Wikipedia language editions from January 2019 to March 2020. RCFilters flags contributions identified by the Objective Revision Evaluation Service (ORES) machine learning system as likely to be damaging [33]. These flags are shown along with existing social signals of quality. We take advantage of a set of arbitrary thresholds built into RCFilters to conduct a quasi-experimental analysis that estimates the causal effect of algorithmic flagging on moderation decisions and that seeks to measure whether algorithmic flags lead to better or worse outcomes for users who are likely to be overscrutinized *ex ante*. Our results suggest that algorithmic flagging can lead to more fair outcomes but that this effect may depend on the specifics of the social signals in question.

Our paper makes several contributions. First, our work answers calls to analyze the impacts of algorithms *in situ* [87, 90, 97] by offering an empirical evaluation of an algorithmic flagging system in an important social computing context. Second, our analysis contributes to an ongoing debate over when and how algorithms might lead to more or less fair outcomes for individuals subject to profiling by human decision makers. Third, our work offers a methodological contribution by presenting a novel quasi-experimental approach that can act as a template for future non-intervention studies of causal effects of algorithmic decision support systems. Finally, our work contributes to social computing system design by suggesting improvements to algorithmic flagging and filtering systems.

2 BACKGROUND

2.1 Moderation in Online Communities

Contemporary online communities are flooded with harassment, spam, misinformation, disinformation, and hate. Users of social media systems frequently and flagrantly violate community and platform rules, various laws, and norms of decency and decorum. Even users acting in good faith can do damage by taking conversations off-topic, undermining the stated purpose of communities, and lowering the quality of discourse or the knowledge goods being produced. Protecting online communities from unwanted activity are content moderators—many of them volunteers—that Gillespie [30] has described as "custodians of the Internet." Moderation work typically involves three tasks: namely, reviewing content or activity, mitigating damage caused by a problematic behavior, and sanctioning users in different ways [30, 49, 52, 86].

Grimmelmann [32] defined moderation as "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." Discussions of content moderation often focus on individuals occupying formal roles as moderators with special rights and responsibilities. For example, several of the moderators in Gillespie's [30] account are professional moderators working for major platforms such as Facebook and Twitter. Several moderators, and nearly all of them on platforms such as Reddit and Discord [49, 52, 69], work as volunteers but occupy similar positions of formal authority and responsibility. That said, the work of moderation is also distributed across regular community members [53, 62]. In Wikipedia, for example, the bulk of moderation activity as defined by Grimmelmann occurs as normal users review, vet and undo the work of others to mitigate damage and sanction users they believe have behaved badly [79].

2.1.1 Sanctions. Sanctioning involves enforcing norms in ways that attempt to discourage future misbehavior. It is a core part of moderation work because it encourages compliance with norms by communicating that rules will be enforced [47, 89]. Although it also serves to mitigate damage, removing content is a common form of sanctioning because it communicates that an action was inappropriate [79]. Halfaker et al. [36] showed that removing content is an effective sanction and results in higher quality contributions by the reverted contributor in Wikipedia. Similarly, Srinivasan et al. [89] found that people whose comments were removed from Reddit were less likely to violate norms in the future.

Although the goal of most sanctioning is to steer participants toward productive behaviors, the effect is often to deter participation. This can be particularly problematic with well-meaning newcomers who often violate norms because they have not yet learned the ropes [1, 34, 36]. Sanctioned newcomers are less likely to continue participating, especially in the absence of clear explanations from moderators [34, 47, 54, 80, 91]. On Wikipedia and similar communities, high rates of sanctioning can help explain declines in participation and may be an obstacle to building a community that includes diverse participants [34, 60, 91].

2.1.2 Meta-norms. No moderation system is perfect. Moderators inevitably make mistakes and apply sanctions in ways that are arbitrary and unfair. This is particularly challenging to avoid in distributed moderation models used on sites such as Slashdot or Wikipedia where moderation is conducted by large and diverse groups of untrained and loosely coordinated users. Sanctions can be particularly demotivating to newcomers who feel that sanctions are unfair and incorrect [30, 47, 89]. Consequently, steps that make moderation more fair might decrease the negative effects of sanctions on community growth.

One way to improve fairness in moderation is through governance structures that enforce accountability [25]. Toward this end, Slashdot famously created tools for "meta-moderation" that allowed all users to evaluate the decisions of moderators [62]. Users whose moderation decisions were controversial or at odds with the opinions of other Slashdot members would not be given moderation privileges again. Although formal systems for meta-moderation remain rare, behaviors that take action against controversial sanctions are common and serve a similar social function [17]. "Meta-norms," which prescribe when and how one should issue sanctions against violations of first-order norms [43] are particularly relevant. Reagle [81] documented the formalization of meta-norms on Wikipedia and Piskorski and Gorbatâi [79] showed how Wikipedia users maintain meta-norms by undoing sanctions in ways that effectively sanction the originally sanctioning user.

2.1.3 Flagging and Algorithmic Triage. Moderators in large online communities can face incredible challenges in scaling their work to handle an enormous mass of content and user activity [30, 52, 85, 86]. In interviews conducted by Kiene et al. [52], small teams of volunteers tasked with maintaining order in large communities described their work as akin to "running a small city." Some

platforms deal with scale by employing more paid moderators. However, the work involved can be exploitative, challenging, traumatizing, and expensive [83]. Volunteer moderator teams frequently find it difficult to identify, train, and integrate new members as they grow [53]. On average, teams become less likely to add new members as their communities grow [88].

For these reasons and others, it is often impossible for communities to scale moderation resources such that human moderators can review all activity. As a result, many moderation systems implement flagging so that a wider group of users can report content for review by moderators [32]. If users reliably flag problematic behaviors, flagging can mitigate issues of scale because moderators focus their attention on behavior that is flagged. Obviously, flagging is far from a perfect solution. From the perspective of a flagged user, flagging can seem arbitrary and opaque [17]. From a moderator perspective, flagging is flawed because disgruntled users can coordinate to overwhelm moderators and target opposing viewpoints [17]. Finally, given that traditional flagging systems continue to rely on volunteer labor, they often fail to fully address issues of scale, leaving many bad actions unflagged, unreviewed, and unsanctioned.

To address this final limitation, communities have turned to algorithmic flagging systems that use computer programs to automatically mark content for review by human moderators [51, 52, 85]. Although some of these systems rely on keywords, regular expressions, or heuristics, more advanced and flexible versions of these systems use predictions from machine learning models. These systems are seen as promising answers to the problem of moderation at scale because they can be easily be used to review an enormous volume of behaviors, they may be less vulnerable to strategic flagging, and they may be more reliable than human reviewers.

Algorithmic flagging systems can be thought of as human-in-the-loop versions of similar computational systems that engage in fullly automated moderation. For example, numerous digital platforms utilize the PhotoDNA system to automatically identify and remove child pornography [30]. Similarly, Wikipedia's ClueBot NG uses a machine learning predictor to automatically remove vandalism [28]. Although they play a critical role in reducing moderation workloads, fully automated systems are uncertain enough in most of their assessments that they are typically only considered useful in defending against the most clear-cut examples of misbehavior [30].

Some machine learning systems that are designed to classify bad behavior are used as a form of algorithmic triage. While the most egregious examples of bad behavior are dealt with by automatic systems, other possible norm violations are flagged for review by human moderators. For example, Reddit allows moderators to define a system of rules based on regular expressions to automatically flag content for further review [48]. Algorithmic flagging systems based on machine learning occupy the vanguard of online activity regulation and numerous examples have been described in recent scholarship. Chandrasekharan et al. [14] described a system for Reddit communities to share information and collaborate on automatic flagging that accounts for differences between rules of different communities. Wulczyn et al. [95] presented a system for classifying harassing behavior on Wikipedia. Finally, Halfaker and Geiger [33] developed the ORES system to predict the quality of contributions and content on Wikipedia.

2.2 Will Algorithmic Flagging Decrease Discrimination Of Overprofiled Users?

One of the most important debates in contemporary technology policy is the degree to which the introduction of algorithms into socially consequential decision making leads to more or less fair outcomes [15, 55, 76, 87]. Much of this debate focuses on arguments about whether algorithms will amplify or entrench discrimination and focuses on biases introduced by training data [4, 12, 84]. Discrimination is the deferential treatment of individuals based on membership in a group. Economists of discrimination distinguish between taste-based and statistical discrimination [6, 8, 78]. Taste-based discrimination is driven by preferences for members of one group and includes both

ideologically-driven racism and implicit bias. Statistical discrimination occurs when social signals—visible and socially salient characteristics, such as group memberships—are instrumental in driving decisions. Statistical discrimination can also lead to unequal outcomes for certain groups.

2.2.1 Social Signals. Although most discussions of discrimination focus on high-stake contexts such as banking, labor markets, and criminal justice, moderation in online communities is also ripe for statistical discrimination based on visible social signals. When interacting in face-to-face groups, people can observe—and discriminate on the basis of—visible signals of status, group membership, psychological states, or cultural identity [21, 77, 82]. Because the invisibility of these signals in online communities creates a barrier to regulation, sociability, and cooperation, communities use devices such as profile images and biographies, avatars, or visualizations of activity as tools for self-presentation and signals of membership [21, 63]. Disclosing information on profiles can provide signals helpful for people using prototypes [31], building social capital [24], and developing trust [67]. Formal reputation systems such as karma on Reddit and Slashdot or badges on StackExchange can be important signals of commitment, quality, and trustworthiness [32, 61, 72]

Even without user profiles or formal reputation systems, participants in online communities use subtle signals to draw conclusions about each other [20, 23, 46]. Sparse cues such as usernames or communication styles can be signals of personality, gender, and identity [21, 37, 41]. Tests of community-specific technical or cultural knowledge can identify newcomers and, similar to formal reputation systems, they may be more challenging to fake than biographical information [7, 21, 32]. In peer production projects, prior contributions can be inspected for information about expertise, work styles, and the future value of a newcomer [68].

In several online communities such as Wikipedia, users can elect to participate anonymously, under more-or-less stable pseudonyms, or using their real names. Masking signals of gender, race, age, (dis)ability, or status can appear to equalize and free individuals from oppressive prejudices and stereotypes [22, 26]. On the other hand, the presence or absence of a stable user identity is itself an essential signal because persistent identities make it possible to build up reputation, social capital, and trust and the inability to do so is associated with misbehavior [31, 42].

2.2.2 Will algorithmic flagging reduce overprofiling? Online community moderators can use social signals to discover and respond to misbehavior, but this can lead to statistical discrimination. Wikipedia's *Missing Manual* advises would-be vandal fighters on Wikipedia to "consider the source" when "estimating the likelihood that an edit is vandalism" [10]. Because newcomers and anonymous users are more likely to violate rules, moderators may rely on social signals of newness to find bad behaviors or to decide if an ambiguous contribution was made in bad faith. Increased scrutiny and skepticism can translate into an increased likelihood of sanction, simply for being new or anonymous. Statistical discrimination emerges because moderators are more likely to scrutinize and sanction new or anonymous contributors who have legitimate reasons for contributing.

Ethical philosophers have objected to the way social signals are used in online moderation activity. Dutch philosopher Paul de Laat adopted the concept of "profiling" from legal scholar Frederick Schauer to argue against the use—and even the public display of—social signals such as registration status and experience levels in the user interfaces used for moderation because they are prone to "overuse" [18, 19]. It should be noted that discriminating by attributes such as newness does not raise the same legal or constitutional concerns as discrimination against protected classes such as race or religion. Online communities establish their own norms and may choose to protect or target certain attributes on the basis of a specific community's values. For example, while discussing Wikipedia, de Laat argues that "overuse" is unethical, immoral, and inconsistent with the community's founding principles of transparency and equality. Drawing on de Laat, we refer to individuals with social signals that elicit undue scrutiny as "overprofiled."

Although an important debate continues over the use of algorithmic predictions in domains like criminal sentencing, proponents of algorithms argue that they could reduce discrimination and inequality [55, 90]. Algorithms can reproduce statistical discrimination, but they might be less biased than the alternative: human decisions that would presumably rely heavily, if perhaps subconsciously, on salient social signals such as race. Critics suggest that algorithms simply obscure this discrimination behind complex mathematical models that are difficult to understand, interrogate, or challenge.

Although this debate is challenging to resolve in the case of criminal justice, algorithmic flagging in online community moderation provides a setting with lower stakes and more detailed data. If we apply arguments proposing that algorithms can reduce discrimination to community moderation, we would conclude that algorithmic triage systems would reduce the impact of discrimination among overprofiled individuals by making misbehavior by all kinds of users visible to moderators. If algorithmic flagging reduces overprofiling bias, then it will have a smaller effect on overprofiled users than on others. If algorithms simply reproduce discrimination, we would find no such difference. This leads us to our first research question: ***[RQ1]*** *How will flagging an action change the likelihood an action is sanctioned for overprofiled editors compared with others?*

Algorithmic fairness researchers use specific criteria to quantify biases encoded in algorithmic predictors and the fairness of resulting decisions [4, 15, 73]. These criteria are often developed for settings where model predictions are equivalent to decisions. For example, Kusner et al. [59] define demographic parity in terms of model predictions, whereas Mitchell et al. [73] define it in terms of human decisions. In algorithmic flagging, decisions are informed by algorithms but left to humans. Therefore, we distinguish between the fairness of predictions and the fairness of decisions and refer to our criteria as "decision system fairness metrics" following Mitchell et al.'s [73] use of the term "decision system."

We first consider demographic parity, as shown in Equation 1, which means that the probability of a decision (D) is statistically independent of a protected attribute (A) [4, 59]:

$$P(\widehat{D}|A = 0) = P(\widehat{D}|A = 1) \tag{1}$$

An algorithmic flagging system will have demographic parity concerning registration status if the probability that an action is flagged is the same for actions by overprofiled and underprofiled editors. Our analysis of RQ1 thus evaluates how flagging shapes demographic parity for sanctioning decisions.

2.3 Will Algorithmic Flagging Increase Fairness?

A system might lack demographic parity by sanctioning one group more than others but still be justifiable if all sanctions are fair. What does it mean for a sanction to be fair? The subject of fairness in algorithmic systems is a major subject of debate in computing and AI. There are several different approaches to conceptualizing fairness, and no algorithmic predictor can satisfy them all [4, 13, 56, 73, 93, 96]. While such approaches focus on discrimination built into machine learning programs, we seek a concept of fairness that reflects the standards of relevant communities of practice. We find one in the concept of "meta-norms" from social psychology and James Coleman's sociological conception of norm maintenance. Drawing from these sources, we define unfair sanctions as those that a community is unwilling to let stand—i.e., sanctions that are themselves the subject of sanction [16, 43, 79]. For example, norms in Wikipedia govern right and wrong ways of editing wiki pages. Sanctions of first-order norm violations are governed by meta-norms about what sorts of contributions merit sanction. Following Piskorski and Gorbatâi [79], we describe a

sanction as *controversial*—i.e., in likely violation of a meta-norm—if it, in turn, is sanctioned by a third community member.

A controversial sanction suggests that the initial edit was not truly damaging (i.e., $D = 1$ but $Y = 0$ where $Y = 1$ means an edit was truly damaging). Thus, a controversial sanction is analogous to false positive classification by a machine predictor ($\widehat{Y} = 1$ but $Y = 0$, where $\widehat{Y} = 1$ means the machine predicts that an edit is damaging). The false positive rate quantifies the amount of unfair treatment a group experiences, but it does not compare unfair treatment between groups. Therefore, is not strictly speaking an algorithmic fairness criterion. However, changes in the false positive rate of the decision system (shown in Equation 2) quantify how flagging is increasing or decreasing the rate of unfair sanctions.

$$P(D = 1|Y = 0, \widehat{Y} = 1) - P(D = 1|Y = 0, \widehat{Y} = 0) \tag{2}$$

Relying on this definition of fairness, our second research question asks how algorithmic flagging shapes the fairness of sanctioning in terms of the rate of sanctions for meta-norm violations: ***[RQ2]*** *How will flagging an action change the chances it receives a controversial sanction?*

Influential theoretical frameworks in social computing seem to predict competing answers to this second question. First, dual-process models of behavioral economics suggest that people will tend to rely on "salient signals" for rapid decision making in conditions of uncertainty and imperfect information [9, 55, 92]. When human moderators use social signals to choose behavior to review or sanction, these attributes serve as salient signals but remain far from perfect signals of quality. Algorithmic flags provide an additional salient signal but are also far from perfect [33]. Indeed, algorithmic flagging systems are typically designed to minimize the risk of missing bad behaviors by surfacing large numbers of false positives (i.e., non-problematic behaviors) and relying on human moderators to make final decisions. Of course, if human moderators use algorithmic flags as salient signals, they may reproduce algorithms' false predictions. In this case, controversial sanctions will increase.

A second perspective suggests that algorithmic flags can increase fairness. Several online communities have institutionalized rules, norms and meta-norms and act as highly bureaucratic organizations [11, 79]. Max Weber described how bureaucratic organizations construct and use two concepts of what he called "rationality:" substantive rationality and formal rationality [94]. Substantive rationality refers to how bureaucratic organizations use policies, routines and hierarchy to define their collective values and goals. Formal rationality refers to the use of calculated decision making, such as that involving productivity or financial metrics, in the pursuit of goals [65]. Following Weber, Kreiss et al. [58] argued that increasing substantive rationality through bureaucratic policies in online communities can lead to more fair outcomes.

Although less explored by scholars of online communities, there are also reasons to believe that increasing formal rationality in moderation decisions might also enhance fairness, at least in online communities with mature normative systems. In such contexts, algorithmic flagging systems can enact formal rationality by estimating the probability and displaying an authoritative signal that an action runs afoul of shared behavioral standards. Adopting algorithmic flaggin can thus mark a shift away from idiosyncratic individual decision-making and toward increasing the use of formalized rationality. Through this lens, an algorithmic flagging system—even one that encodes biases—can be a "carrier of formal rationality" [65], leading to governance that is more in line with community meta-norms and to a decrease in controversial sanctions.

Next, we consider how changes in the false positive rate of the decision system depends on overprofiling. This corresponds to evaluating decision system fairness in terms of equality of

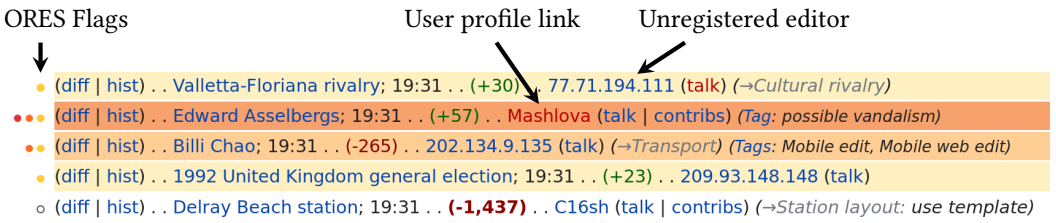


Fig. 1. Screenshot of Wikipedia edit metadata on Special:RecentChanges with RCFilters enabled. Highlighted edits with a colored circle to the left side of other metadata are flagged by ORES. Different circles and highlight colors (white, yellow, orange and red in the figure) correspond to different levels of confidence that the edit is damaging. Users can configure which colors are shown. Visible social signals include registration status (i.e., whether a username or an IP address is shown) and whether an editor's user page and user talk page exist. RCFilters does not specifically flag edits by new accounts, but does support filtering changes by newcomers.

opportunity (shown in Equation 3) [39, 73]:

$$P(D = 1|Y = 0, A = 0) = P(D = 1|Y = 0, A = 1) \tag{3}$$

Equality of opportunity is satisfied when the false positive rate of a decision system does not depend on the protected attribute. Equality of opportunity for registration status would mean that registered and unregistered editors that make good edits have equal chances of having their contributions accepted.

Our third research question asks whether algorithmic flagging systems will increase or decrease equality of opportunity: ***[RQ3]*** *Within the set of sanctioned actions, how will the effect of flagging an action on controversial sanctions depend on whether contributors are overprofiled?*

Once again, influential theoretical frameworks in social computing research seem to point in opposite directions. Under dual-process psychological models, both social signals and algorithmic flags might cue moderators to issue sanctions and might substitute for one another. In this case, we would hypothesize that flagging would have a more positive effect on controversial sanctions among underprofiled contributors, who had previously been relatively ignored, than it does among the overprofiled individuals, who were always scrutinized. Conversely, if the larger effect of algorithmic flagging is helping moderators comply with meta-norms, it simply will not matter whether contributors are overprofiled.

3 EMPIRICAL SETTING

We aim to answer our three research questions through a field evaluation of an algorithmic flagging system called RCFilters, which was deployed on 23 different Wikipedia language editions between January 2019 and March 2020. RCFilters stands for "Recent Changes filters." The term "Recent Changes" refers to a page on Wikipedia that allows viewers to see the most recent changes made to the site.[1] As Figure 1 shows, RCFilters adds a set of flags represented as colored dots on the left side of the list of recent contributions. Social signals are also visible, including registration status and whether a user has created a profile page. Although dense with information regarding recent edits and hyperlinks, the page is immediately understandable to Wikipedia moderators. When deployed, the RCFilters interface appears both on "Recent Changes" as well as on "watchlists"—a

[1]For example, the Recent Changes page for English Wikipedia is available here: https://en.wikipedia.org/wiki/Special:RecentChanges (Archived: https://perma.cc/BNZ3-E9D5)

special version of "Recent Changes" that shows only edits to the subset of pages that a user has elected to follow. RCFilters must be enabled by each user on their Wikipedia user preferences page.

Algorithmic flagging in the RCFilters system is powered by the ORES edit quality models trained to predict whether edits are labeled "damaging" or "not damaging." The models are gradient boosted decision trees trained on a mixture of human-labeled Wikipedia edits and edits made by established editors that are assumed to be "not damaging."

It should be noted that ORES models do not merely reproduce profiling patterns typical of moderation on Wikipedia. The interface for labeling training data obscures social signals from the volunteer Wikipedians doing labeling work and its models are predictive of damage from users that are not anonymous or newcomers. Nevertheless, as discussed in §8, ORES encodes biases against unregistered editors and—to a lesser extent—against editors without user pages. ORES was designed neither to merely support quality control in Wikipedia, nor to optimize precision, recall, or fairness but to enact Wikipedian principles of openness, transparency, and community accountability—to "deploy efficient machine learning at scale for content moderation … in ways that enable volunteers to develop and deploy advanced technologies on their own terms" [33]. More information on the philosophy, design and implementation of ORES can be found in Halfaker and Geiger [33].

4 METHODS

Our analysis is based on a regression discontinuity design (RDD) that aims to estimate causal the effects of flagging by RCFilters on moderator behavior in Wikipedia [44, 45, 64]. Common in empirical economics, RDDs are quasi-experimental in that they resemble a randomized control trial for data points in the neighborhood of an arbitrary cutoff [45, 64]. RDDs model how an outcome depends on this cutoff and a continuous "forcing variable." The idea behind an RDD is that observations immediately below and above the cutoff will be equal in expectation after adjusting for any underlying (i.e., "secular") trend. For example, RDDs used in econometrics might estimate the effect of passing a test by comparing the outcomes of people who barely passed and failed. One benefit of an RDD over a field experiment based on A/B tests is that it can provide ecological validity and support causal claims without subjecting users to intervention without consent [5, 50]. Although they remain rare in computing, RDDs have been used in recent publications in social computing [42, 75].

Our forcing variable is the score from the ORES machine learning system. Our cut-off variables are a set of arbitrarily chosen operating points used by RCFilters. Our outcomes are constructed by creating two variables that indicate whether a revision's author is overprofiled as well as variables that indicate whether each revision was reverted or subject to a controversial revert. We discuss each in turn before introducing our analytic approach.

4.1 Data and Measures

We build our dataset from two publicly available tables of Wikimedia history published by the Wikimedia Foundation (WMF).[2] Although Wikipedia is published and collaborated on in several languages, the vast majority of knowledge regarding collaboration on Wikipedia is derived from studies of English Wikipedia [38, 40]. To support generalizability, we analyze data from 23 language editions of Wikipedia where edit quality flags are displayed in the RCFilters interface. To ensure that we have variation in our outcomes, we exclude wikis with less than three edits above and below each threshold (see §4.1.1) from each sub-analysis. For all of our analyses, our unit of analysis

[2] https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Edits/Mediawiki_history (Archived: https://perma.cc/CPM6-PY6F; https://dumps.wikimedia.org/other/mediawiki_history/readme.html (Archived: https://perma.cc/3DDJ-9FXS)

is the *revision*. Revisions correspond to a single edit to a page by a participant on Wikipedia. We exclude revisions by bots since we care about how algorithmic flagging and social signals are used by human moderators. Following guidance for RDDs [64], we include only revisions very near to RCFilters thresholds, with ORES scores within 0.03 of the thresholds.

To manage the total size of our dataset, we analyze a sample that we construct by stratifying along several dimensions: Wikipedia language edition; user registration status (§4.1.4); whether the editor has a user page or not (§4.1.4); whether an edit was reverted in 2 hours, 48 hours, or 30 days; and whether the revert was controversial (§4.1.3). Then, we sample 5000 edits from within unique combinations of the variables. If there are less than 5000 edits in the given strata, we include all of them. We adjust for this stratification using sample weights throughout our analysis. Since RCFilters was introduced to different wikis at different times, we sample edits during the period immediately following the introduction of ORES but weight our sample according to the number of edits to each wiki over the entire study period. The numbers of observations sampled at each threshold, from each Wiki, and for each model are available in the supplementary material.

4.1.1 ORES scores and RCfilter thresholds. The continuous forcing variable used in our RDD analysis is a score from the ORES algorithm described in §3. Scores range from 0 to 1 and reflect the predicted probability that a revision is damaging. Because the ORES system has been under continual development over time, we obtain ORES scores created at the times revisions were made from a log maintained by the WMF. The treatments in our analysis are whether edits to Wikipedia are flagged by RCFilters. These flags are applied if, and only if, a score from ORES exceeds a threshold. This use of thresholds at arbitrary operating points is a feature of most algorithmic flagging systems. The intuition behind our RDD is that—after adjusting for small differences in quality associated with marginally higher or lower scores—edits with ORES scores immediately above and below an arbitrary threshold will be similarly likely to receive both first-order and controversial sanctions. Consequently, any discontinuous change in reverts at one of the thresholds used by RCFilters can be attributed to the flag.

RCFilters uses multiple thresholds corresponding to green, yellow, orange, and red flags. By default, only orange, and red flags are shown, but users can configure which colors to display. Green flags and filters are to help Wikipedia editors find good edits. Our analysis considers only red, orange, and yellow flags, which correspond to thresholds making different trade-offs between precision (the proportion of flagged edits that are truly damaging) and recall (the proportion of truly damaging edits that are flagged). The red flag is labeled "very likely damaging" and corresponds to a high precision threshold. Orange flags corresponds to a "likely damaging" label with greater recall but less precision. Edits with a yellow flag are "maybe damaging" with a high recall but lower precision. RCFilters' thresholds are truly arbitrary and have changed over time and across language editions in response to shifts in the precision and recall of ORES models and in response to community feedback. We were able to collect data on threshold configuration, fully trained ORES models, code, and the precise time that changes were deployed in the WMF server admin log. We combined these data to identify the precise thresholds that were active for each revision in our dataset.

4.1.2 Sanctions. Our outcome variable for answering RQ1 must capture sanctioning in Wikipedia. Following a large body of other social computing research, we measure sanctions as identity reverts [e.g., 34, 36, 79, 91]. Identity reverts occur when a user undoes another user's edit by restoring a page to an earlier state and are measured by comparing hashes of page revisions [36].

That said, identity reverts are an imperfect measure of sanctioning. It is also possible for an individual to "self-revert" by undoing their own edit. We therefore only treat a revision as reverted if it was undone, but not by a self-revert. We also limit our measure of sanctioning to revisions that

are undone within 48 hour to avoid problems related to mass revert actions such as "blanking" of pages that result in false positives. We are confident that 48 hours is a reasonable window because most damage to Wikipedia will be undone within that amount of time [28] and a 48 hours window will include reverts caused by RCFilters since any effect of RCFilters is likely to occur quickly.

4.1.3 Controversial sanctions. Our outcome variable for answering RQ2 and RQ3 measures controversial sanctions. We follow Piskorski and Gorbatâi [79] by measuring controversial sanctions as identity reverts that are subsequently reverted by a third party. Specifically, we label a sanction as controversial if the sanction is undone by a third editor who was not the original editor or the reverting editor. Such interactions likely correspond to cases in which a third party observes the initial revert, disagrees with the initial sanction and then acts to reverse the sanction.

4.1.4 Social signals. Answering our RQ1 and RQ3 requires that we identify underprofiled and overprofiled individuals in our empirical setting. Drawing from research and documentation for Wikipedia moderators, we identify two such measures shown in the RCFilters interface shown in Figure 1. Our first measure is whether an editor was logged into an account. Unregistered editors act on Wikipedia without logging in and registered contributors are those that edit with accounts. Because they are identified by their IP address rather than by a chosen username, unregistered editors are also referred to as "IP editors" or "anons." Unregistered editors are associated with misbehavior and have long had a controversial status on Wikipedia [71]. Geiger and Ribes described how tools for moderators highlight unregistered editors [29]. de Laat argued that unregistered editors on Wikipedia are overprofiled in that they are at higher risk to have their contributions rejected unfairly [18, 19].

Second, the RCFilters interface indicates whether the editor has created a user page. User pages are Wikipedia's version of profile pages. Not having a user page is a social signal of newness because most committed users will create a user page early into their experience in Wikipedia [3]. The presence or absence of pages in Wikipedia is indicated with a subtle user interface clue: links to pages that do not exist are rendered in red, whereas links to pages that exist are blue. For example, Figure 1 shows the user "Mashlova" whose name is shown in red and would be identified as a newcomer. de Laat cited the absence of a user page as a second example of an indicator of vandalism that will result in overprofiling [19]. We measure whether an editor's user page exists at the time of a given contribution by matching the titles of user pages against the editor's username and checking if the creation of the user page was prior to the edit in question. We only include registered editors in our analysis of overprofiling based on user pages.

5 ANALYTIC PLAN

Our analysis comprises Bayesian logistic regression models in two parallel analyses. The first analysis treats our dichotomous measure of whether edits are reverted as an outcome. This begins with an "adoption check" (§6) that describes the causal effects of flagging on reverts in general. The adoption check is a prerequisite to answering our research questions. The rest of the first analysis (§7.1) answers RQ1 by comparing the effect of RCFilters on edits by overprofiled users to its effect on other editors. Our second analysis is very similar but uses controversial reverts as the outcome, and analyzes only reverted edits to model the probability that a revert is controversial. It begins by answering RQ2 (§7.2) in an analysis similar to the adoption check but with controversial sanctions as an outcome and with a dataset limited to overprofiled users. The rest of the second analysis (§7.2) answers RQ3 and is similar to RQ1 but with controversial reverts as the outcome in place of reverts.

Although our models use different sets of edits and outcomes, they all have the same logistic regression structure shown in Equation 4.

Threshold	Edit type	N.	Prop.
Maybe dam.	Not reverted	12,403,717	0.87
Maybe dam.	Rev. controversial	69,395	0.00
Maybe dam.	Rev. not cont.	1,757,866	0.12
Maybe dam.	**Total**	14,230,978	1.00
Likely dam.	Not reverted	1,254,219	0.55
Likely dam.	Rev. controversial	31,652	0.01
Likely dam.	Rev. not cont.	1,009,108	0.44
Likely dam.	**Total**	2,294,979	1.00
V. likely dam.	Not reverted	58,474	0.15
V. likely dam.	Rev. controversial	12,545	0.03
V. likely dam.	Rev. not cont.	323,762	0.82
V. likely dam.	**Total**	394,781	1.00

(a) Counts and proportions of edits by whether an edit was reverted or controversially reverted in the neighborhood of each threshold.

Threshold	Editor type	N.	Prop.
Maybe dam.	Reg. No User Page	4,006,466	0.28
Maybe dam.	Reg. User Page	3,797,451	0.27
Maybe dam.	Unregistered	6,415,271	0.45
Maybe dam.	**Total**	14,219,188	1.00
Likely dam.	Reg. No User Page	281,964	0.12
Likely dam.	Reg. User Page	26,459	0.01
Likely dam.	Unregistered	1,982,985	0.87
Likely dam.	**Total**	2,291,408	1.00
V. likely dam.	Reg. No User Page	21,630	0.05
V. likely dam.	Reg. User Page	687	0.00
V. likely dam.	Unregistered	371,499	0.94
V. likely dam.	**Total**	393,816	1.00

(b) Counts and proportions of edits by whether an editor was registered or had a user page in the neighborhood of each threshold.

Table 1. Summary statistics from our full dataset.

$$\log\left(\frac{P\left(Y_r\right)}{1-P\left(Y_r\right)}\right) = \alpha_1\left(score_r - c_{jw}\right) + \tau_j \mathbf{1}\left[score_r > c_{jw}\right] + \alpha_2\left(score_r - c_{jw}\right)\mathbf{1}\left[score_r > c_{jw}\right] + \alpha_w \tag{4}$$

Our goal is to estimate τ_j which is the causal effect of being flagged at level j, where $j \in \{1, 2, 3\}$ corresponds to labels of "maybe damaging," "likely damaging" and "very likely damaging." For each cutoff on each wiki, we select revisions whose ORES scores are within a ±0.03 window of the cutoff (c_{jw}). Following established approaches to RDD, we fit "kink" models that allow for a change in slope at the discontinuity [64, 66]. The slope before the discontinuity is α_1 and the change in slope is α_2. The indicator function is represented by $\mathbf{1}$. Our models include fixed effects for wiki (α_w) to account for differences in the rates of sanctioning between wikis.

We use Bayesian inference to estimate our models for two reasons. First, virtually all edits above the "very damaging" level are reverted in some of the wikis we analyze. The presence of near-perfect "separation" creates estimation problems for classical numerical approaches [2]. Preferred solutions to this problem in non-Bayesian frameworks include penalized likelihood methods that introduce bias. Our Bayesian approach uses weakly-informative priors that are conservative but avoid the problem of separation. The second reason we use Bayesian inference is that it makes it easy to compare estimates across models. Prior work at CSCW by Gan et al. [27] used a similar rationale for adopting Bayesian logistic regression. In Bayesian analysis, fitted models take the form of posterior distributions constituting a probability distribution of model coefficients conditional on our model, data and priors. We consider a hypothesis supported if it is consistent with at least 95% of posterior draws. In other words, we accept a given hypothesis if our parameter estimate has the predicted sign and the 95% credible interval does not contain 0. This is the Bayesian analog to testing a hypothesis with $\alpha = 0.05$. We fit our models using the rstanarm package (version 2.19.3) and the default priors that are provided for reference in the supplementary material.

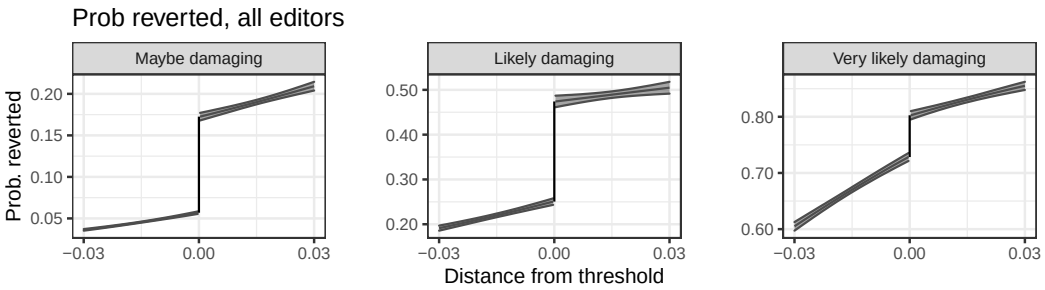


Fig. 1. Marginal effects plot showing model predicted relationship between ORES score and the probability that an edit will be reverted around the cutoffs for all contributors with 95% credible intervals.

6 ADOPTION CHECK

Before presenting results from hypothesis tests associated with our research questions, we first establish that RCFilters was adopted by Wikipedia moderators and that it had an effect on sanctioning behavior. This establishes a baseline necessary to answer RQ1 regarding the differential effects of RCFilters between overprofiled users and others. Null effects in RQ1 might simply reflect that the system was not used. A successful adoption check rules out this possibility and sets up a credible null hypothesis test for RQ1.

We test the hypothesis that flagging increases the probability that an edit is reverted to demonstrate that RCFilters flags are being used by Wikipedia moderators. Our estimates for τ_j—as described in §5—should be positive if Wikipedia moderators are using flags in RCFilters to review potentially damaging edits.

We find strong evidence that RCFilters was adopted and impacted sanctioning. Figure 1 visualizes this evidence: a marginal effects plot that illustrates our models' predicted likelihood of reverts across different ORES scores in the neighborhood of the thresholds. In each such plot, the x-axis shows the distance from the threshold such that discontinuities at 0 represent the effect of being flagged. The plots show modeled values for the English language edition of Wikipedia but are representative of relationships across all wikis.[3] Figure 1 shows discontinuous increases in the likelihood of reversion at the "maybe damaging" and "likely damaging" thresholds in the left and center panels. We find the greatest effect at the "maybe damaging" threshold ($\tau_1 = 1.23$ [1.19; 1.28]).[4] The effect at the "very likely damaging" threshold shown in the right-most panel is smaller ($\tau_3 = 0.41$, [0.35; 0.46]).

The impacts of the "maybe damaging" and "likely damaging" flags on the likelihood of sanctioning are enormous. Figure 1 shows that likelihood of a revert for an edit just below the "maybe damaging" threshold is between 5.5% and 5.8%, indicating that reverts of unflagged edits are relatively rare. Being flagged with the "maybe damaging" flag causes a dramatic increase in the reversion probability to between 16.8% and 17.7% for edits just above the threshold. The effect of algorithmic flags at the "likely damaging" level is even more stark. We estimate that edits just below the "likely damaging" threshold are likely to be reverted between 24.3% and 25.8% of the time, whereas similar edits just above the threshold are reverted between 46.1% and 48.7% of the time. Being flagged at the "very

[3]Because intercepts are the only part of our model that depend on Wikis, slopes and the discontinuities caused by algorithmic flagging represent our inference over all our data.

[4]All τ parameter estimates are reported as log-odds ratios. The bracket notation indicates the 95% credible interval. In other words, the most likely value of the parameter is 1.23, but there is a 95% probability that the parameter lies in the interval [1.19; 1.28].

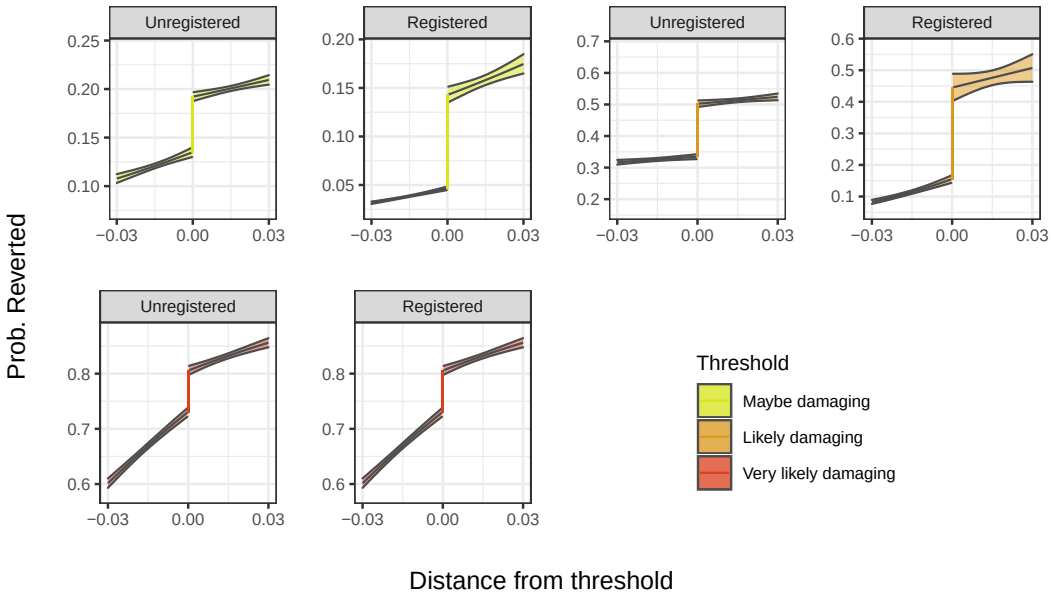


Fig. 2. Results for RQ1 comparing unregistered and registered contributors are displayed in a marginal effects plot showing the model predicted relationship with 95% credible intervals between ORES scores and reverts around the thresholds that trigger flags.

likely damaging" threshold causes an increase in reversion probability from between 72.1% and 73.5% to between 79.5% and 81%.

7 RESULTS

7.1 RQ1: Effect of Flagging on Sanctioning

In our first research question (RQ1), we seek to understand how the increase in sanctioning caused by flagging affects discrimination against overprofiled users. If algorithmic flagging reduces overprofiling, as some computer scientists have argued [55], the effect of flagging will be more scrutiny on users who are more likely to be given a pass. If algorithms simply reproduce discrimination, we will find no difference. Results for hypothesis tests answering this question are shown in Figure 3, which visualizes the point estimates and credible intervals for differences in the causal effects of flagging on reverts between unregistered and registered contributors and between contributors with and without user pages. Values greater than 0 indicate that our estimated effect for the other users is greater than that for the overprofiled group.

In support of the idea that algorithmic flagging can reduce overprofiling bias, we find that the overall effect of flagging is to increase demographic parity between registered and unregistered editors. Aggregating our posteriors over all three thresholds shows that the average effect over the three thresholds is greater for registered editors than for unregistered editors ($\frac{1}{3}\sum_{j=1}^{3}\tau_j^{\text{Reg}} - \tau_j^{\text{Unreg}} = 0.45\ [0.16; 0.6]$). The effect of flagging on reverts of registered editors is greater than the effect for unregistered editors at both the "maybe damaging" threshold ($\tau_1^{\text{Unreg}} - \tau_1^{\text{Reg}} = 0.8\ [0.71; 0.89]$) and the "likely damaging" threshold ($\tau_2^{\text{Unreg}} - \tau_2^{\text{Reg}} = 0.78\ [0.58; 0.97]$). For an action by an unregistered contributor near to the "maybe damaging" threshold, being flagged increases the odds of being

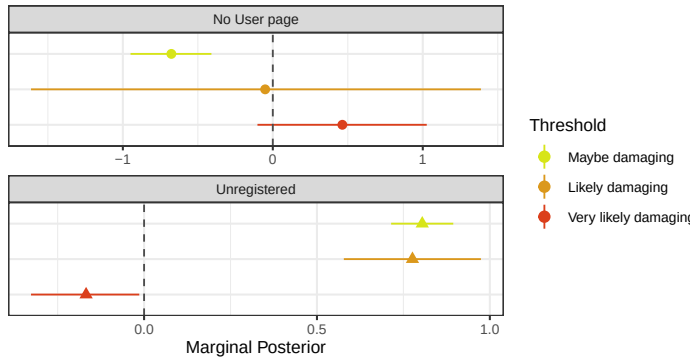


Fig. 3. Results for RQ1 showing point estimates and 95% credible intervals for differences in the causal effect of flagging on sanctioning between overprofiled contributors and others. A value greater than 0 indicates that our estimates of the effect for underprofiled contributors are greater than those for overprofiled contributors.

reverted by a factor of between 1.45 and 1.6 times. This is significantly less than the increase of between 3.16 and 3.68 times for registered contributors.

However, at the "very likely damaging" threshold we find that the effects of flagging are stronger for unregistered editors than for registered editors ($\tau_2^{\text{Reg}} - \tau_2^{\text{Unreg}} = -0.17$ $[-0.33; -0.01]$). Being flagged increases the odds that an action is reverted by a factor of between 1.43 and 1.62 times for an unregistered editor and by 1.11 and 1.49 times for registered contributors. However, as Table 1 shows, a far greater number of actions receive scores near to lower thresholds. Thus, we focus on the lower thresholds in the following discussion.

Figure 2 lets us interpret our models by making it possible to visually compare the effects of being flagged between overprofiled and underprofiled editors at a given threshold because the y-axes in each row span an identical range. The top-left panel shows how our models' linear predictions of how the probability of sanctioning for unregistered contributors at the "maybe damaging" threshold jumps between 4.8 and 6.7 percentage points, from 13.5% to 19.2% on average. For registered editors, shown in the top-right of Figure 2, we estimate a jump of between 9.1 and 10.3 percentage points, from 4.6% to 14.3% on average. This is between 3.3 and 4.6 percentage points greater than the jump for unregistered editors. For unflagged edits that ORES scores near the "maybe damaging" threshold, an unflagged unregistered contributor has about the same odds of being sanctioned as a flagged registered contributor.

The bottom row of Figure 2 shows that the change in sanctioning probability at the "likely damaging" threshold is between 9.5 and 15.2 percentage points greater for registered editors than for unregistered editors. For unregistered contributors, shown in the bottom-left of Figure 2, being flagged as "likely damaging" increases the probability of revert between 15 and 18.6 percentage points, from 33.5% to 50.2% on average. But for registered editors, shown in the bottom-right of Figure 2, we detect an even bigger jump of between 23.7 and 34.6 percentage points, from 15.5% to 44.5% on average. For actions that ORES scores near the "likely damaging" threshold, unflagged actions by unregistered editors are far more likely to be reverted. Once flagged, actions by registered and unregistered editors are reverted at relatively similar rates.

These results show that flagging causes an increase in a decision system's demographic parity concerning registration status. Actions by unregistered contributors that fall just above the cutoffs are much more likely to be reverted due to RCFilters—but the gap between actions by registered and unregistered contributors is much smaller when RCFilters has flagged an edit as "maybe

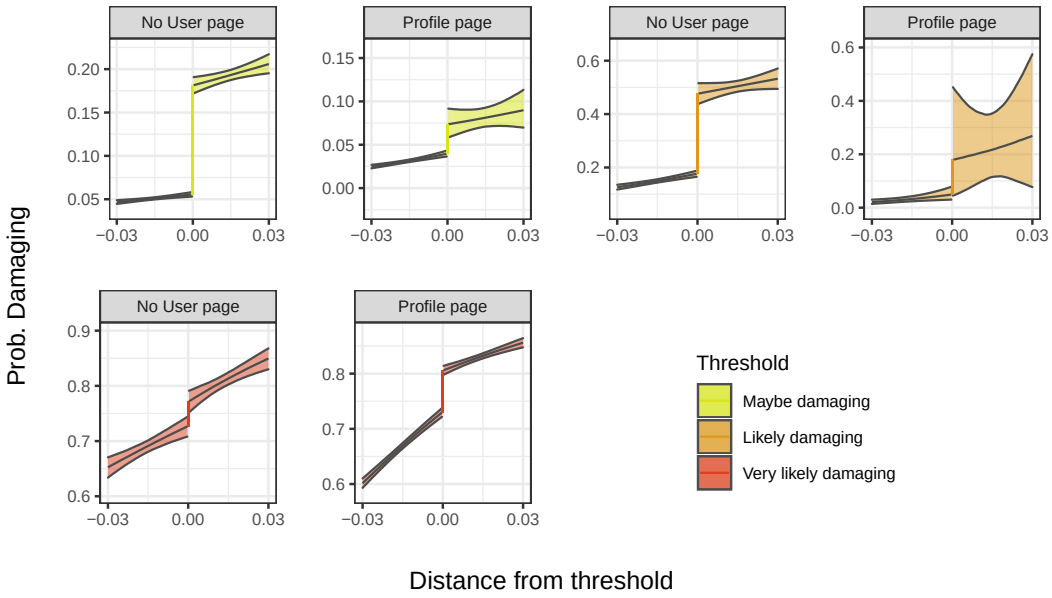


Fig. 4. Results for RQ1 comparing contributors with and without user pages. Each panel shows a marginal effects plot with 95% credible intervals of the modeled relationship between ORES scores and reverts around the thresholds that trigger flags.

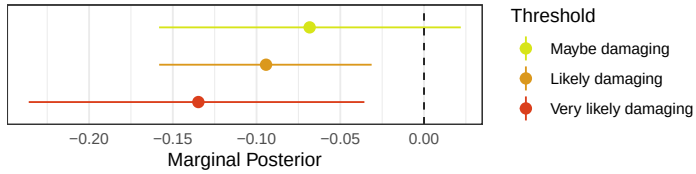
damaging" or "likely damaging." In this way, our analysis suggests that algorithmic flagging can reduce overprofiling bias.

Surprisingly, our results for our second measure of over-profiling in Wikipedia suggest a dynamic that is opposite in sign to the differences we observe between registered and unregistered editors at the "maybe damaging" threshold ($\tau_1^{\text{NoUP}} - \tau_1^{\text{UP}} = -0.68$ $[-0.95; -0.41]$). At the "likely damaging" ($\tau_2^{\text{NoUP}} - \tau_2^{\text{UP}} = -0.05$ $[-1.61; 1.39]$) and the "very likely damaging" ($\tau_2^{\text{NoUP}} - \tau_2^{\text{UP}} = 0.46$ $[-0.1; 1.03]$) thresholds we do not detect differences in effect size between contributors with and without user pages. At the "maybe damaging" threshold, we find that flagging increases the odds that an editor without a user page is reverted between 3.47 and 4.06 times. This is significantly more than the increase of between 1.47 and 2.46 times for registered contributors.

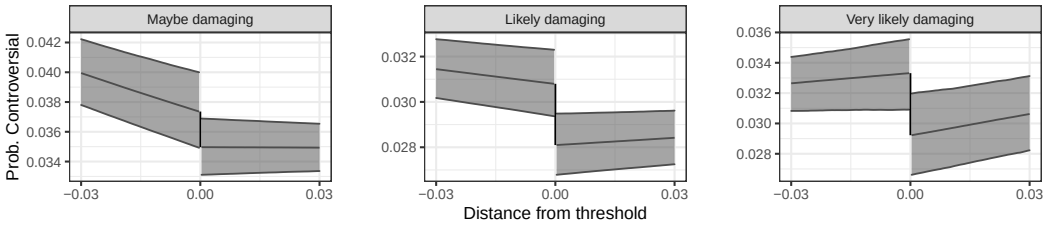
As above, we interpret these odds ratios using marginal effects plots shown in Figure 4. The top-left plot in the figure shows our models' linear predictions of the probability of reverting for contributors without user pages near to the "maybe damaging" threshold. For these editors, being flagged as "maybe damaging" increases the chances of sanctioning by 11.4 and 13.8 percentage points, from 5.6% to 18.1% on average. In the top-right of Figure 4, we see a jump of between 2.2 and 4.8 percentage points, from 4% to 7.4% on average for editors that have created user pages. This is between 9.7 and 8.4 percentage points less than the jump for contributors without user pages.

7.2 RQ2: Effect of flagging on controversial sanctioning

Consistent with the idea that algorithmic flagging can support fairness, we find that having an ORES score cross the "likely damaging" or "very likely damaging" thresholds decreases the chances that a revert will be controversial for unregistered editors. These results are visualized in Figure 5a. We have less confidence in the effect at the "maybe damaging" threshold because our 95% credible interval includes 0 ($\tau_1^{\text{Unreg}} = -0.07; \text{CI} = [-0.16; 0.02]$).

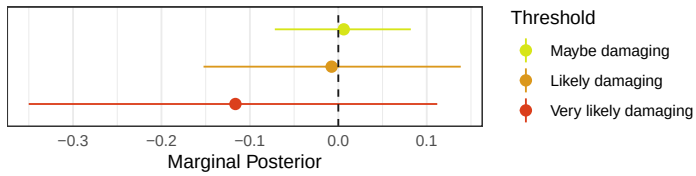


(a) Parameter estimates and 95% credible intervals for the effects of flagging on whether reverts are controversial for unregistered editors.

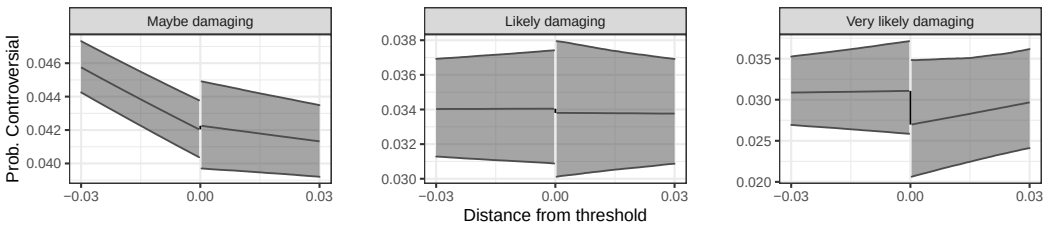


(b) Marginal effects plots with 95% credible intervals for models predicting whether a revert is controversial, for unregistered editors.

Fig. 5. Results for RQ2: flagging causes a small but detectable decrease in the likelihood that an action by an unregistered contributor receives a controversial sanction.



(a) Parameter estimates and 95% credible intervals for effects of flagging on whether reverts are controversial for editors without user pages.



(b) Marginal effects plots with 95% credible intervals for models predicting whether a revert is controversial, for contributors without user pages.

Fig. 6. Results for RQ2 comparing contributors with user pages to those without show no detectable effect of flagging on controversial sanctioning.

We estimate that being flagged at the "likely damaging" level results in a change in the odds that a sanction is controversial by a factor between 0.85 and 0.97. Figure 5b shows the modeled relationship between ORES scores and the probability of a controversial sanction in the neighborhood of the thresholds for English Wikipedia. On the left plot, we see that being flagged changes unregistered contributor's likelihood of a controversial revert from a possible increase of 0.27 percentage points to a possible decrease of 0.55 percentage points, a change from 3.08% to 2.81% on average.

We observe a similar effect of flagging at the "very likely damaging" threshold ($\tau_2^{\text{Unreg}} = -0.13; \text{CI} = [-0.24; -0.04]$): the odds that a revert is controversial are between 0.79 and 0.97 times smaller. On the right side of Figure 5b, we find that being flagged decreases the probability that a sanction to an action by an unregistered editor is controversial by between 0.11 and 0.89 percentage points, a change from from 3.33% to 2.92% on average.

However, we did not detect effects of flagging when the reverted editor lacks a user page at the "maybe damaging" ($\tau_1^{\text{NoUP}} = 0.01; \text{CI} = [-0.07; 0.08]$), "likely damaging" ($\tau_2^{\text{NoUP}} = -0.01; \text{CI} = [-0.15; 0.14]$), or "very likely damaging" ($\tau_3^{\text{NoUP}} = -0.12; \text{CI} = [-0.35; 0.11]$) thresholds. Our results for RQ2 for unregistered editors show that flagging decreases the rate of controversial sanctions. Although controversial sanctions do not precisely correspond to false-positive sanctions, we take this finding as evidence that flagging decreases the false positive rate of the decision system. We address the inconsistencies between our results for unregistered editors and editors without user pages in our discussion (§9).

7.3 RQ3: Social signals and effects of flagging on controversial sanctioning

To answer RQ3, we largely replicate the analysis conducted for RQ1 with the dependent variable used in RQ2. Results shown in Figure 7 provide weak evidence that a decrease in controversial sanctioning may be greater for registered than for unregistered contributors at the "maybe damaging" ($\tau_1^{\text{Reg}} - \tau_1^{\text{Unreg}} = 0.04\ [-0.06; 0.14]$), "likely damaging" ($\tau_2^{\text{Reg}} - \tau_2^{\text{Unreg}} = 0.07\ [-0.05; 0.2]$), and "very likely damaging" ($\tau_3^{\text{Reg}} - \tau_3^{\text{Unreg}} = 0.02\ [-0.23; 0.27]$) thresholds. However, our evidence weakly suggests that the effect for contributors with user profiles is greater than those for without at the "maybe damaging" threshold ($\tau_1^{\text{UP}} - \tau_1^{\text{NoUP}} = 0.05\ [-0.08; 0.17]$) but the opposite seems true at the "likely damaging" threshold ($\tau_2^{\text{UP}} - \tau_2^{\text{NoUP}} = -0.26\ [-0.79; 0.26]$) and "very likely damaging" ($\tau_3^{\text{UP}} - \tau_3^{\text{NoUP}} = -0.16\ [-0.9; 0.56]$) thresholds. None of these estimates are statistically significant at the 95% level.

8 THREATS TO VALIDITY

Our results are subject to a range of threats to validity that pertain to our ability to make causal claims, rule out alternative explanations, and establish the generalizability of our findings. First, there are several threats to our ability to draw causal inferences that are common to RDDs. Formally, RDDs model an outcome Y as a function of a continuous "forcing variable" Z, other covariates, and a cutoff c such that $Z > c$ determines treatment assignment. In principle, treatment assignment conditional on Z is "as good as random" under two assumptions: (1) that agents have at most limited control over $Z > c$, and (2) that the relationship between Y and Z is smooth [64]. Although the assumptions required for causal inference are fundamentally unverifiable, we believe that our RDD provides relatively strong evidence of causal relationships between flagging and sanctioning.

Our treatment, being flagged in RCFilters, is an ideal candidate for an RDD from the perspective of assumption (1) because editors are unlikely to have much control over the scores that their edits receive. Although attempts to evade sanction by specially crafting edits to evade algorithmic detection are hypothetically possible, the authors of ORES and RCFilters believe they are unrealistic and very unlikely to be widespread. Assumption (2) would be violated if any unobserved treatments

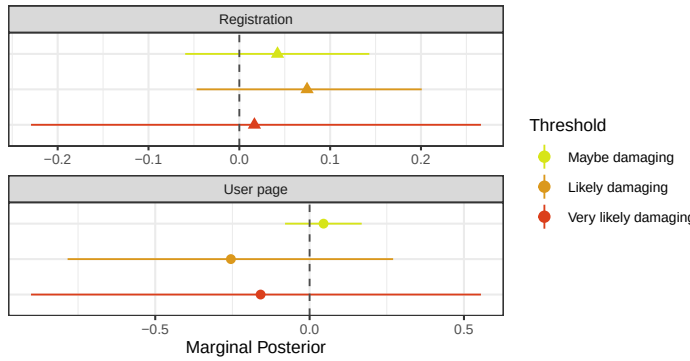


Fig. 7. Results for RQ3 showing the difference in our parameter estimates between overprofiled editors and others with 95% credible intervals. Values greater than 0 would indicate that the effect for underprofiled editors is greater than that for overprofiled editors.

affect our outcomes at discrete levels of ORES scores. This is certainly possible because ORES makes scores available via a public API. Indeed, we are aware of bots that automatically revert edits triggered by the "very damaging" threshold on some of the Wikipedia language editions in our sample and therefore have more reason to doubt results at this threshold. Despite this threat, our conclusions regarding how algorithmic flagging shapes fairness are substantively similar whether we consider this threshold or not. Although we identified one anti-vandalism tool—a system called Huggle discussed in §9—that collects ORES damaging scores, it uses ORES scores as one feature in its own algorithmic model and, by default, presents predictions from this model to users as a list of edits sorted in order of likelihood of vandalism. Given these facts, we believe that it is unlikely that Huggle users will drive discontinuities in the relationship between ORES scores and our outcomes.

A limitation of RDD analysis is that it estimates effects for observations in the neighborhood of the cutoff and results may not generalize far away from the cutoff. Compared with most RDD analysis, ours has the advantage of multiple different thresholds. Although our results for the "likely damaging" and "maybe damaging" thresholds are substantively similar, causal effects may diverge more at operating points we have not considered. Future work on algorithmic bias using RDD should consider that results may depend on the choice of operating points used as RDD cutoffs.

An additional threat to validity is raised by the extent to which the ORES models encode biases concerning editors who are unregistered or without profile pages. To assess this threat, we analyzed the bias of ORES models for each wiki that had deployed the system on December 19th 2020 using their human-labeled training data according to the *conditional calibration* approach to evaluating model bias [73].[5] In our case, this involves comparing the rate of damaging edits predicted by the model to the true rates for each type of editor. We find that ORES exhibits bias against both unregistered editors and editors without user pages but that the extent of bias against unregistered editors is much greater. These findings are opposite in sign to what we would expect if model bias were driving our results. We present detailed results from this analysis in our online supplement.

Our study design is also limited in that we cannot present causal evidence of the impact of social signals. Although RCFilters's algorithmic flags are distributed in a quasi-experimental way, overprofiled status is not. There are a range of possible systematic differences between overprofiled users and others that might be driving our results for RQ1 and RQ3. For example, if damaging

[5]We chose conditional calibration as our fairness metric because it does not depend on the choice of threshold. This simplifies the analysis of a decision system with multiple thresholds.

edits by contributors who are unregistered or lack user pages are more difficult for ORES to detect, that might drive our findings of a decrease in overprofiling for RQ1. Although we believe that this particular threat is unlikely because it would require that overprofiled contributors be systematically more sophisticated than others—something our experience with ORES suggests is unlikely—we cannot rule out either the specific threat or a range of other possibilities. A promising direction for future work might involve experiments or quasi-experiments that can jointly vary social signals and algorithmic flagging.

Additionally, system designers will likely want to know how overall rates of sanctioning and controversial sanctions change before and after a system such as RCFilters is launched. Unfortunately, our analysis cannot answer this question directly. In preliminary work, we attempted to draw a statistical comparison between Wikipedia governance before and after the introduction of ORES but high temporal variation in sanctioning behaviors made this type of aggregate change difficult to measure. Future studies should organize with communities to conduct planned and principled field experiments to study the causal effects of introducing such systems in online communities using the model being pioneered by Matias and Mou [70].

Finally, a set of largely unanswerable threats involves questions of generalizability across our measures and empirical contexts. Although our theory of interactions between algorithmic flags and social signals is general, and although we study RCFilters across 23 distinct communities, languages, and cultures, we study a single moderator tool on one platform. We cannot claim that our findings generalize beyond the specific pool of communities that we study. Additionally, we have considered only a small subset of possible social signals that may be used in online community moderation. Clearly, we also cannot claim that our settings are representative of moderation in online communities in general. Like most other empirical studies in social computing, we must sadly leave these questions for further research.

9 DISCUSSION

In the broadest strokes, our work is potentially good news for advocates of algorithmic flagging in social computing systems. It provides some evidence supporting the idea that algorithmic flagging can reduce discrimination in the form of overprofiling bias and that it can increase fairness. Our adoption check (§6) provides strong evidence that RCFilters drives behavior and our answers to RQ1 (§7.1) suggests that flagging can level the playing field by increasing decision system demographic parity between unregistered and registered Wikipedia editors. Flagged edits by these contributors are reverted at similar rates, but unflagged edits of comparable quality by registered editors are reverted relatively infrequently. More good news comes in the form of our answer to RQ2 (§7.2) that suggests that flagging is associated with a decrease in controversial sanctions among some overprofiled users and provides evidence that algorithmic flagging systems can help moderators more accurately issue sanctions.

When it comes to the details, however, the picture that emerges from our results is much more contingent and mixed. Our analysis used two different measures of overprofiling in Wikipedia but the pattern of our results diverged substantially between the two. The optimistic story about the effects of algorithmic flagging on overprofiled users only describes our results for unregistered Wikipedia users. Our evidence on overprofiled users without user pages is much weaker and points, in part, in the direction of algorithmic flagging increasing discrimination. Why do these results diverge? What do these divergent results mean for theory?

One possible explanation is that editors without user pages are, quite simply, not particularly overprofiled. Of the two social signals we consider, registration status attracts far more attention from academics and community members in discussions of Wikipedia vandalism [e.g., 42]. Our analysis for RQ2, where we did not detect changes in controversial sanctions for editors without

user pages, is also consistent with the notion that contributors without user pages may not be overprofiled. If algorithmic flagging systems help moderators more accurately issue sanctions by reducing overprofiling, then flagging would not decrease controversial sanctioning for editors that are not overprofiled. However, this alone does not explain why the effect for editors without profile pages was larger than for editors with them.

Our results might be explained if model bias against contributors without user pages means that the set of flagged edits from these users are less damaging than flagged edits by contributors who have profile pages. As discussed in §8 and documented in our online supplement, ORES models are sometimes biased against contributors without user pages, but they are even more biased against anonymous contributors. Our results make sense if the overprofiling of anonymous editors outweighs model bias against them, but the reverse is true for editors without user pages.

It is also plausible that our mixed results are evidence that algorithmic flags will substitute for some social signals used in overprofiling while reinforcing others. Our study analyzes only two of many possible social signals that online community moderators might use. A better understanding of which signals drive sanctioning misbehavior can help explain if and when algorithmic triage systems can increase fairness. Our results suggest that algorithmic flags can substitute for some social signals and reduce overprofiling in online community moderation. Our results also suggest that they might reinforce social signals, make overprofiling worse, or introduce new forms of unfairness through encoded bias. Unfortunately, outcomes resulting from myriad factors acting at once are likely contingent on details of sociotechnical arrangements and difficult to know *ex ante*.

Although RQ2 suggests that algorithmic flagging can increase fairness for overprofiled contributors, our null results for RQ3 mean that we could not detect a difference in this effect between overprofiled editors and others. Uncertainty in our models for RQ3 is high enough that parameter values consistent with a substantive average effect that is either positive or negative are plausible. A null effect for RQ3 might also be explained if meta-norms and improved information are more important to controversial sanctioning than bias introduced by algorithmic flags or social signals acting as cues.

Our work has several important implications for designers of algorithmic flagging systems and sociotechnical systems. Scholars of human computer interaction, science and technology studies, and the law have all called for analyses of algorithmic fairness to move beyond biases inherent in algorithms to consider the systemic and downstream effects of algorithms in use [87, 90, 97]. Ultimately, we recommend that operators of algorithmic flagging systems should continuously evaluate decision system fairness metrics and seek to improve them according to their values. In that the ORES model is, itself, biased against overprofiled users, our results suggest that evaluating the fairness of model predictions is only one piece of understanding how an algorithmic system shapes fairness in contexts such as online community moderation.

Future work should rigorously construct and critique decision system fairness criteria in terms of their consequences. The algorithmic fairness literature often treats algorithmic predictions as equivalent to final decisions. Our work shows that sociotechnical decision systems with humans in the loop face distinctive and contextually sensitive epistemic, ontological, and ethical questions about how decision system fairness should be defined or measured [55, 87].

Decision system fairness is particularly important in open production communities such as Wikipedia because of the trade-offs between quality control and the essential tasks of supporting newcomers and encouraging contribution [34, 74]. Past work has shown that increased quality control efforts correspond to a decrease in newcomer engagement and have hypothesized that one mechanism is increased scrutiny of newcomers [34, 91]. Similarly, although blocking anonymous edits to wikis has shown been shown to cause a decrease in reverted edits, it also leads to a decrease in positive contributions [42]. While it may be intuitive to think about edits that get sanctioned as

obvious vandalism, many of the edits flagged at the "maybe damaging" threshold are authored by well-meaning newcomers [34]. There's a potentially high cost to sanctioning these low quality but well-intentioned contributions. We believe that our results point to the benefit of tracking changes in the rate of sanctions to sensitive groups of community members in order to assure that such well-meaning contributors are not being driven away.

There are also lessons to learned from the impressive degree with which RCFilters shapes behavior. Although the choice of operating points in algorithmic systems is often framed as purely about trading off precision and recall, our work demonstrates that these choices can have a range of other important consequences. Our disparate findings at the "very likely damaging" threshold for overprofiling based on registration status reveal that an algorithmic tool might improve fairness at a given operating point but decrease it at another. Although thresholds allowed us to explore the effects of flagging on sanctioning behavior, this arbitrary flagging of actions applied by RCFilters brought disproportionate attention to contributions just above the thresholds compared to contributions just below. Designers should think about whether using thresholds to trigger flagging in moderation interfaces is a fair practice at all. Our results show that this leads to sanctioning behavior that is, like the thresholds, arbitrary.

What types of designs might support quality control support models that scrutinize contributions in proportion to the likelihood that the contributions deserve to be sanctioned? We see some inspiration in Huggle, a counter-vandalism tool for Wikipedia which sorts actions by the likelihood that they are damaging.[6] Huggle users are encouraged to review the highest likelihood edits first and only move onto lower likelihood edits once those reviews are complete. Such a user experience might increase efficiency and fairness by better concentrating moderator attention wherever it can have the greatest benefits.

10 CONCLUSION

As algorithmic flagging becomes more integrated into online community moderation, it is important to understand its effects and consequences on overprofiling and fairness. We use a regression discontinuity analysis of the RCFilters to find and sanction misbehavior by volunteers on Wikipedia to consider how the use of algorithmic flagging and social signals interact. We find that by drawing moderator attention to misbehavior by registered participants, algorithmic flagging can reduce overprofiling in certain contexts. We also find that algorithmic flagging can support fairness by decreasing controversial sanctions of unregistered contributors. Our results also suggest that the same system may have much less effect, and might even increase discrimination, for other types of overprofiled users.

Studies of machine learning in high-stake settings like employment, education, and criminal justice trace how algorithms can encode discriminatory patterns in human behavior but might also improve fairness compared with human biases. Although the stakes are much lower, such questions are also pertinent to the use of machine predictions for online community moderation. We find that tools for predictive governance in a sociotechnical system can reduce overprofiling but their effects are also difficult to anticipate.

Although our analysis of overprofiling based on registration status supports a rosy account of algorithmic flagging, our analysis of overprofiling based on user pages does not. While contributors without user pages may be less overprofiled compared to unregistered contributors, our results also suggest that the interaction between algorithmic flagging and social signals is complex and contingent. We suggest a need for future work that describes the kinds of social signals that are used in practice and explains how different types of information may be used alongside algorithmic flags.

[6] See discussion in [35]

Finally, we present a methodological approach that we hope future studies of algorithmic tools in real-world sociotechnical systems might build upon to establish the causal effects of algorithmic systems without experimental intervention.

ACKNOWLEDGMENTS

We are grateful to the anonymous CSCW reviewers and associate chairs for their keen insights and feedback. We would also like to thank the Wikimedia Foundation for its support, members of the WMF analytics team including Andrew Otto, Luca Toscano, and Joal Allemandou for help with data access and computing infrastructure and members of the WMF research team including Jonathan Morgan and Miriam Redi for feedback early in project development. Thanks also go to members of the Community Data Science Collective who provided multiple rounds of feedback and contributed to copyediting including Kaylea Champion, Charles Kiene, Stefania Druga, Sohyeon Hwang, Jeremy Foote, and Aaron Shaw. We also thank the WMF staff and volunteers who developed the systems we analyze including Roan Kattouw, the main developer of RCFilters, and the developers of ORES including Amir Sarabadani and Andy Craze. Special thanks to Amanda TeBlunthuis. Finally we owe an extra special thanks to the Wikipedia contributors whose digital traces we analyze. Portions of this work were facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington. Financial support for this work came from the Wikimedia Foundation, from the National Science Foundation graduate research fellowship program #2016220885, and from the University of Washington.

DATA ACCESS

A replication dataset including ORES scores, thresholds, and our sample of Wikipedia revisions, along with all of our code has been placed in the Harvard Dataverse archive and is available at the following URL: https://doi.org/10.7910/DVN/E0RYJ4

REFERENCES

[1] B. Thomas Adler and Luca de Alfaro. 2007. A Content-Driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. Association for Computing Machinery, Banff, Alberta, Canada, 261–270.

[2] Paul Allison. 2004. Convergence Problems in Logistic Regression. In *Numerical Issues in Statistical Computing for the Social Scientist*. John Wiley & Sons, Ltd, 238–252.

[3] Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works*. No Starch Press.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness in Machine Learning*. fairmlbook.org.

[5] Solon Barocas and Helen Nissenbaum. 2015. Big Data's End Run around Anonymity and Consent. In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Julia I Lane (Ed.). Cambridge University Press, New York, NY.

[6] Gary Stanley Becker. 1957. *The Economics of Discrimination*. University of Chicago Press, Chicago.

[7] Michael Scott Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Fifth International AAAI Conference on Weblogs and Social Media*. AAAI Publications, Palo Alto, CA.

[8] Marianne Bertrand and Esther Duflo. 2016. *Field Experiments on Discrimination*. Technical Report w22014. National Bureau of Economic Research, Cambridge, MA.

[9] Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2012. Salience Theory of Choice Under Risk. *The Quarterly Journal of Economics* 127, 3 (Aug. 2012), 1243–1285.

[10] John Broughton. 2008. *Wikipedia the Missing Manual*. Pogue Press/O'Reilly, Beijing; Sebastopol, CA.

[11] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't Look Now, but We've Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 1101–1110.

[12] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI Now 2017 Report. *AI Now Institute at New York University* (2017).

[13] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–15.

[14] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-Based System to Assist Reddit Moderators. *Proc. ACM Hum.-Comput. Interact. CSCW* 3 (Nov. 2019), 174:1–174:30.

[15] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163.

[16] James S. Coleman. 1988. Social Capital in the Creation of Human Capital. *Amer. J. Sociology* 94 (1988), S95–S120.

[17] Kate Crawford and Tarleton Gillespie. 2016. What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society* 18, 3 (March 2016), 410–428.

[18] Paul B. de Laat. 2015. The Use of Software Tools and Autonomous Bots against Vandalism: Eroding Wikipedia's Moral Order? *Ethics and Information Technology* 17, 3 (Sept. 2015), 175–188.

[19] Paul B. de Laat. 2016. Profiling Vandalism in Wikipedia: A Schauerian Approach to Justification. *Ethics and Information Technology* 18, 2 (June 2016), 131–148.

[20] Judith Donath. 2007. Signals in Social Supernets. *Journal of Computer-Mediated Communication* 13, 1 (Oct. 2007), 231–251.

[21] Judith Donath. 2014. *The Social Machine: Designs for Living Online.*

[22] Vitaly J. Dubrovsky, Sara Kiesler, and Beheruz N. Sethna. 1991. The Equalization Phenomenon: Status Effects in Computer-Mediated and Face-to-Face Decision-Making Groups. *Human–Computer Interaction* 6, 2 (June 1991), 119–146.

[23] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. 2006. Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication* 11, 2 (Jan. 2006), 415–441.

[24] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. 2011. Connection Strategies: Social Capital Implications of Facebook-Enabled Communication Practices. *New Media & Society* 13, 6 (Sept. 2011), 873–892.

[25] Seth Frey, P. M. Krafft, and Brian C. Keegan. 2019. "This Place Does What It Was Built for": Designing Digital Institutions for Participatory Change. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 32:1–32:31. arXiv:1902.08728

[26] Eric J. Friedman and Paul Resnick. 2001. The Social Cost of Cheap Pseudonyms. *Journal of Economics & Management Strategy* 10, 2 (2001), 173–199.

[27] Emilia F. Gan, Benjamin Mako Hill, and Sayamindu Dasgupta. 2018. Gender, Feedback, and Learners' Decisions to Share Their Creative Computing Projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 54:1–54:23.

[28] R. Stuart Geiger and Aaron Halfaker. 2013. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?. In *Proceedings of the 9th International Symposium on Open Collaboration (OpenSym '13)*. ACM, New York, NY, 6:1–6:6.

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

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

[31] Sonja Grabner-Kräuter and Sofie Bitter. 2015. Trust in Online Social Networks: A Multifaceted Perspective. *Forum for Social Economics* 44, 1 (Jan. 2015), 48–68.

[32] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17 (2015), 42–109.

[33] Aaron Halfaker and R Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. 4, 148 (Oct. 2020), 37.

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

[35] Aaron Halfaker, R. Stuart Geiger, and Loren G. Terveen. 2014. Snuggle: Designing for Efficient Socialization and Ideological Critique. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 311–320.

[36] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. ACM, New York, NY, 163–172.

[37] Jeffery T. Hancock and Phillip J. Dunham. 2001. Impression Formation in Computer-Mediated Communication Revisited: An Analysis of the Breadth and Intensity of Impressions. *Communication Research* 28, 3 (June 2001), 325–347.

[38] Noriko Hara, Pnina Shachaf, and Khe Foon Hew. 2010. Cross-Cultural Analysis of the Wikipedia Community. *Journal of the American Society for Information Science and Technology* 61, 10 (2010), 2097–2108.

[39] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]* (Oct. 2016). arXiv:1610.02413 [cs]

[40] Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 291–300.

[41] Susan C. Herring. 2000. Gender Differences in CMC: Findings and Implications. *Computer Professionals for Social Responsibility Journal* 18, 1 (2000), 0.

[42] Benjamin Mako Hill and Aaron Shaw. 2020. The Hidden Costs of Requiring Accounts: Quasi-Experimental Evidence from Peer Production. *Communication Research* (2020), 30.

[43] Christine Horne. 2001. The Enforcement of Norms: Group Cohesion and Meta-Norms. *Social Psychology Quarterly* 64, 3 (2001), 253–266.

[44] Guido W. Imbens and Thomas Lemieux. 2008. Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics* 142, 2 (Feb. 2008), 615–635.

[45] Robin Tepper Jacob, Pei Zhu, Marie-Andrée Somers, and Howard Bloom. 2012. A Practical Guide to Regression Discontinuity. *MDRC Working Papers on Research Methodology* (2012).

[46] David Jacobson. 1999. Impression Formation in Cyberspace: Online Expectations and Offline Experiences in Text-Based Virtual Communities. *Journal of Computer-Mediated Communication* 5, 1 (Sept. 1999).

[47] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 192:1–192:33.

[48] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5 (July 2019), 31:1–31:35.

[49] Jialun "Aaron" Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-Based Online Communities on Discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 23.

[50] Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen, and Turo Uskali. 2016. Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics. *Media and Communication* 4, 4 (Oct. 2016), 75–85.

[51] Charles Kiene and Benjamin Mako Hill. 2020. Who Uses Bots? A Statistical Analysis of Bot Usage in Moderation Teams. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–8.

[52] 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 (Nov. 2019), 44:1–44:23.

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

[54] Sara E. Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating Behavior in Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design*, Robert E. Kraut and Paul Resnick (Eds.). The MIT Press.

[55] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (Feb. 2018), 237–293.

[56] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]* (Sept. 2016). arXiv:1609.05807 [cs, stat]

[57] Robert E. Kraut, Paul Resnick, and Sara Kiesler. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. MIT Press, Cambridge, MA.

[58] Daniel Kreiss, Megan Finn, and Fred Turner. 2011. The Limits of Peer Production: Some Reminders from Max Weber for the Network Society. *New Media & Society* 13, 2 (March 2011), 243–259.

[59] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4066–4076.

[60] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. 2011. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*. ACM, New York, NY, 1–10.

[61] Cliff Lampe. 2012. The Role of Reputation Systems in Managing Online Communities. In *The Reputation Society*, Hassan Masum and Mark Tovey (Eds.). The MIT Press.

[62] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, Vienna, Austria, 543–550.

[63] Cliff A.C. Lampe, Nicole Ellison, and Charles Steinfield. 2007. A Familiar Face(Book): Profile Elements as Signals in an Online Social Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. ACM Press, San Jose, California, USA, 435–444.

[64] David S. Lee and Thomas Lemieux. 2010. Regression Discontinuity Designs in Economics. *Journal of Economic Literature* 48, 2 (2010), 281–355.

[65] Dirk Lindebaum, Mikko Vesa, and Frank den Hond. 2019. Insights From "The Machine Stops" to Better Understand Rational Assumptions in Algorithmic Decision Making and Its Implications for Organizations. *Academy of Management Review* 45, 1 (May 2019), 247–263.

[66] Stephan Litschig and Kevin M. Morrison. 2013. The Impact of Intergovernmental Transfers on Education Outcomes and Poverty Reduction. *American Economic Journal: Applied Economics* 5, 4 (Oct. 2013), 206–240.

[67] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2397–2409.

[68] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression Formation in Online Peer Production: Activity Traces and Personal Profiles in Github. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 117–128.

[69] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, 1138–1151.

[70] J. Nathan Matias and Merry Mou. 2018. Civilservant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 9:1–9:13.

[71] Nora McDonald, Benjamin Mako Hill, Rachel Greenstadt, and Andrea Forte. 2019. Privacy, Anonymity, and Perceived Risk in Open Collaboration: A Study of Service Providers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–12.

[72] Arpit Merchant, Daksh Shah, Gurpreet Singh Bhatia, Anurag Ghosh, and Ponnurangam Kumaraguru. 2019. Signals Matter: Understanding Popularity and Impact of Users on Stack Overflow. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, San Francisco, CA, USA, 3086–3092.

[73] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2020. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867 [stat]* (April 2020). arXiv:1811.07867 [stat]

[74] Jonathan T. Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and Sympathy: Crafting Positive New User Experiences on Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 839–848.

[75] Sneha Narayan, Nathan TeBlunthuis, Wm Salt Hale, Benjamin Mako Hill, and Aaron Shaw. 2019. All Talk: How Increasing Interpersonal Communication on Wikis May Not Enhance Productivity. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 101:1–101:19.

[76] Cathy O'Neil. 2018. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books, London.

[77] Alex Pentland. 2008. *Honest Signals How They Shape Our World*. The MIT Press.

[78] Edmund S. Phelps. 1972. The Statistical Theory of Racism and Sexism. *The American Economic Review* 62, 4 (1972), 659–661.

[79] Mikołaj Jan Piskorski and Andreea D. Gorbatâi. 2017. Testing Coleman's Social-Norm Enforcement Mechanism: Evidence from Wikipedia. *Amer. J. Sociology* 122, 4 (2017), 1183–1222.

[80] Martin Potthast, Benno Stein, and Robert Gerling. 2008. Automatic Vandalism Detection in Wikipedia. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer, Berlin, Heidelberg, 663–668.

[81] Joseph M. Reagle. 2010. "Be Nice": Wikipedia Norms for Supportive Communication. *New Review of Hypermedia and Multimedia* 16, 1-2 (April 2010), 161–180.

[82] Cecilia L Ridgeway. 2019. *Status: Why Is It Everywhere? Why Does It Matter?*

[83] Sarah Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications* (Jan. 2016).

[84] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678.

[85] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 111–125.

[86] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* 21, 7 (July 2019), 1417–1443.

[87] 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 (FAT* '19)*. ACM, New York, NY, USA, 59–68.

[88] Aaron Shaw and Benjamin Mako Hill. 2014. Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *Journal of Communication* 64, 2 (2014), 215–238.

[89] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content Removal As a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 163:1–163:21.

[90] Megan T. Stevenson. 2017. Assessing Risk Assessment in Action. *SSRN Electronic Journal* (2017).

[91] Nathan TeBlunthuis, Aaron Shaw, and Benjamin Mako Hill. 2018. Revisiting "The Rise and Decline" in a Population of Peer Production Projects. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, 355:1–355:7.

[92] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (Sept. 1974), 1124–1131.

[93] Hanna Wallach. 2019. Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency. *Medium* (Jan. 2019).

[94] Max Weber. 1978. *Economy and Society*. University of California Press, Berkeley, CA.

[95] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17*. ACM Press, Perth, Australia, 1391–1399.

[96] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–12.

[97] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 194:1–194:23.