Algorithmic flags and identity-based signals in online community moderation

Nathan TeBlunthuis, University of Washington
Benjamin Mako Hill, University of Washington
Aaron Halfaker, Wikimedia Foundation

Moderators of online communities and social media platforms review an often large quantity of user generated content and actions to address violations of norms and rules. Upon finding a problematic action, they decide how to respond and whether to sanction the misbehavior. Due to the *problem of scale* moderators may direct their attention according to identity-based signals of individual quality such as reputation, experience, or registration status instead of reviewing every action [Gillespie 2018; Kiesler et al. 2012]. Increasingly, communities and platforms adopt *algorithmic triage* systems to help moderators find actions likely to require intervention [Chandrasekharan et al. 2019]. With growing attention to problems of disinformation and hate speech online, commercial platforms are expanding their pools of paid human moderators, but the work of paid moderators can be exploitative, difficult, traumatizing, and expensive [Roberts 2016]. Moderation is stressful work involving a large number of judgment calls, often ambiguous, that must be made quickly.

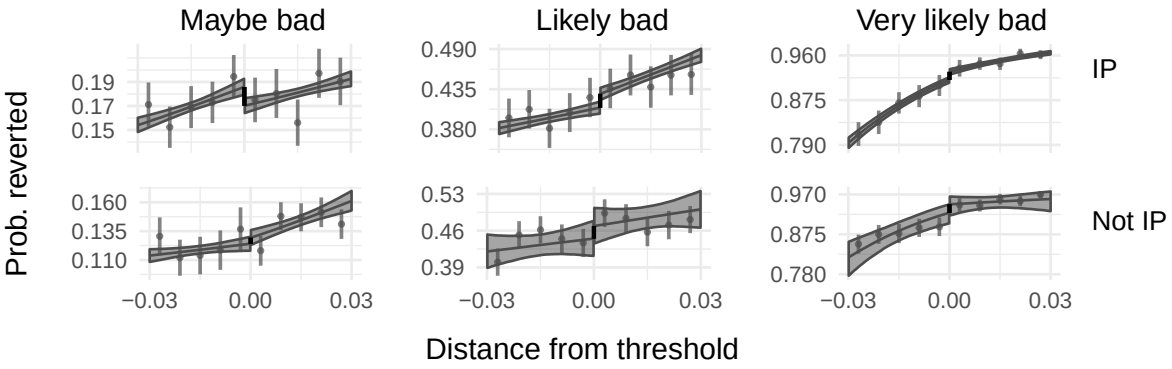


Fig. 1. Marginal effects plot for H1, showing our model's predictions of sanctioning behavior around cutoffs that cause actions to be flagged to moderators. Points with error bars show proportions of edits reverted in 10 bins with 95% confidence intervals.

Drawing from legal philosopher Frederick Schauer's notion of *profiling*, ethicist Paul de Laat argues that characteristics like reputation and registration status become prone to *overuse* by moderators who may concentrate their attention on the activities of a narrow range of users [de Laat 2016; 2015]. Instead of reviewing every action or searching randomly, moderators may direct their attention according to *identity-based signals*: characteristics such as reputation, experience, or registration status associated with quality. But reliance on such signals may lead to *over-profiling* if moderators concentrate their attention on the activities on a narrow range of users.

Over-profiling may lead to *statistical discrimination* when the characteristics correlated with performance or deviance change how decisions are made [Bertrand and Duflo 2016]. Discrimination is when

individuals are treated differently by authorities depending on their group memberships or identity. Economists of discrimination distinguish between taste-based and statistical discrimination [Bertrand and Duflo 2016]. *Taste-based discrimination* is driven by preferences for members of one group or identity and includes both ideological racism and implicit bias. In contrast, statistical discrimination happens when identity-based signals are instrumental to improving the quality of decisions.

Increasingly, online platforms adopt algorithmic triage to predict which content is damaging and surface it to human moderators [Gillespie 2018]. Can algorithmic triage reduce statistical discrimination and reliance on identity-based signals in community and platform governance? Advocates of algorithmic risk prediction in criminal justice settings argue that algorithmic predictions can improve upon the discriminatory and inaccurate decisions of human judges [Kleinberg et al. 2018b]. Yet when moderators or judges can observe identity-based signals, they may still use them in decision making. We hypothesize that algorithmic predictions will have less influence on outcomes for over-profiled individuals compared to under-profiled ones. In other words, our theory is that flagging an action by an algorithm will cause a greater increase in the likelihood of sanction for under-profiled individuals.

We propose that:

H1: Flagging an action causes a greater increase in the likelihood the action is sanctioned when the action is made by an under-profiled individual than when it is made by an over-profiled individual.

We also consider how identity-based signals shape the consistency of sanctioning. When moderators use aspects of user identity such as account age, registration, experience or reputation to choose what contributions to review or whether to sanction behavior, these attributes act as *salient signals*: visible signs used in fast decision making. When faced with many choices where the correct decision is uncertain or where finding and analyzing the information necessary to arrive at a correct decision is difficult, people tend to rely on salient signals instead of alternative information that is more accurate, but difficult to use [Bordalo et al. 2012; Kleinberg et al. 2018a; Tversky and Kahneman 1974].

We propose that algorithmic flags function as a salient signal and therefore that moderators may be more likely to issue controversial sanctions against flagged actions. When an action is flagged, a moderator will be suspicious of it and act conservatively to sanction even if the decision is uncertain because the flag signals to the moderator that the action is problematic. We hypothesize that the increase in sanctioning caused by flagging an action will also lead to an increase in the proportion of sanctions that are controversial.

H2: Within the set of sanctioned actions, flagging an action causes an increase in the likelihood that it receives a controversial sanction.

Finally, we propose that, as with algorithmic flags, identity-based signals function as salient signals that can lead to controversial sanctioning. Similar to **H1**, we hypothesize that using algorithmic flagging alongside identity-based signals will partly, but not entirely, reduce reliance on identity-based signals. Actions by under-profiled individuals will be moderated more conservatively when they are flagged, but more liberally when not flagged. Yet actions by over-profiled individuals will still be moderated conservatively when not flagged. This implies that the increase in controversial sanctions among flagged actions will be smaller for over-profiled individuals compared to under-profiled individuals.

H3: Within the set of sanctioned actions, flagging an action causes a greater increase in the likelihood that the sanction is controversial when the action is by an under-profiled individuals than when it is by an over-profiled individual.

We use a regression discontinuity design (RDD) to estimate the causal effect of being flagged on moderation actions and test our hypotheses by comparing our estimates for over-profiled and under-profiled classes of editors. Given some assumptions, RDDs resemble a randomized control trial for data near to a discontinuity. RDDs model an outcome Y, as a function of a continuous forcing variable Z, other covariates X, and a cutoff c such that $Z > c$ determines treatment assignment. The goal is to

estimate τ, which can be interpreted as the local average treatment effect in the neighborhood of c. We use logistic regression models fit with rstanarm and weakly informative priors that shrink our estimates slightly towards 0.

We analyze data on moderator behavior from several language editions of Wikipedia that have adopted the ORES algorithm for edit quality prediction and the RCfilters flagging and filtering user-interface that it powers. This system flags edits at three different levels ("maybe bad", "likely bad", "very likely bad") when the ORES model's score (our forcing variable) exceeds arbitrary thresholds. The moderation interfaces present information about group memberships associated with damaging behavior, specifically whether an edit is attributed to an IP address or not. Our outcome for **H1** is whether an edit is *identity reverted*, a measure of sanctioning commonly used in Wikipedia research and our outcome for **H2** and **H3** is whether a revert is un-reverted by a third party.

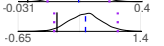
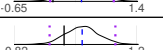
Fig. 2. Screenshot of Wikipedia edit metadata on Special:RecentChanges with RCfilters enabled. Highlighted edits with a colored circle to the left other metadata are flagged by ORES. Different circle colors (yellow and orange in the figure) correspond to different levels of confidence that the edit is damaging.

Table I shows marginal posteriors for the effects of algorithmic flagging on reversion for each editor class and the difference in the estimates between editor classes. Figure 1 shows marginal effects plots for the relationship between ORES scores and the probability of reversion around each threshold. We find support for **H1**, but given the low number edits with low ORES scores made by registered users with in the sample, there is about a 25% chance that the effect for non-IP editors is no greater than for IP editors. Our models predict that an IP edit that scores right below the threshold has 1.2 times the odds of being reverted as an edit that scores right above the thresholds compared to an odds ratio of 1.6 for non-IP editors. We are working on results using a larger sample for the conference.

We tentatively conclude that Wikipedia moderators continued using IP-attribution as a sign of dubious quality as algorithmic flags have a stronger effect for non-IP edits than for IP edits. This supports the notion that moderators in peer production communities like Wikipedia over-profile based on visible characteristics of contributors, but that introducing algorithmic triage systems can reduce statistical discrimination.

We find little support for **H2** or **H3**, that flagging increases controversial sanctioning or that any such increase falls disproportionately on over-profiled editors. Our results from these hypotheses is limited in power by the relative scarcity of controversial sanctions made against registered editors.

Table I. Partial results from RDD analysis showing estimated causal effect of flagging on sanctioning behavior for IP editors, and non-IP editors. The effect is probably greater for non-IP editors compared to IP-editors. Marginal Posterior plots show the distributions of coefficients in our posterior samples. Solid black lines indicate the position of 0, blue dashed lines indicate the mean, and dotted purple lines indicate the boundaries of the 95% credible intervals.

Coefficient	Mean	SD	2.5%	25%	50%	75%	97.5%	Marginal Posterior
τ^{IP}	0.18	0.06	0.07	0.14	0.18	0.22	0.29	
$\tau^{\text{not IP}}$	0.48	0.27	−0.05	0.29	0.48	0.65	1.02	
$\tau^{\text{not IP}} - \tau^{\text{IP}}$	0.30	0.28	−0.24	0.11	0.30	0.48	0.85	

REFERENCES

Marianne Bertrand and Esther Duflo. 2016. *Field Experiments on Discrimination*. Technical Report w22014. National Bureau of Economic Research, Cambridge, MA. w22014 pages. DOI:http://dx.doi.org/10.3386/w22014

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. DOI:http://dx.doi.org/10.1093/qje/qjs018

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. DOI:http://dx.doi.org/10.1145/3359276

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. DOI:http://dx.doi.org/10.1007/s10676-015-9366-9

Paul B. de Laat. 2016. Profiling Vandalism in Wikipedia: A Schauerian Approach to Justification. *Ethics and Information Technology* 18, 2 (June 2016), 131–148. DOI:http://dx.doi.org/10.1007/s10676-016-9399-8

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven. OCLC: on1005113962.

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.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018a. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (Feb. 2018), 237–293. DOI:http://dx.doi.org/10.1093/qje/qjx032

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018b. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (Dec. 2018), 113–174. DOI:http://dx.doi.org/10.1093/jla/laz001

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

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (Sept. 1974), 1124–1131. DOI:http://dx.doi.org/10.1126/science.185.4157.1124