Effects of algorithmic flagging on fairness:
Quasi-experimental evidence from Wikipedia

Nathan TeBlunthuis nathante@uw.edu
Benjamin Mako Hill
Aaron Halfaker

October 11, 2021

University of Washington
Department of Communication
Wikimedia Foundation
Microsoft
The Problem of Scale in Online Moderation

Wikipedia’s administrative tools are often likened to a janitor’s mop. There’s a large amount of potentially damaging activity. Monitoring and correcting misbehavior is expensive. Moderators (often poorly paid workers or volunteers) are the “custodians of the Internet.”
Wikipedia is great. But it’s got issues:

• Hard to attract and retain contributors

• Hard to attract and retain contributors—especially the diverse and marginalized.

• These problems may drive “knowledge gaps.”

• Wikipedia’s quality control system is involved.

Figure 2. The English Wikipedia’s editor decline. The number of active, registered editors (>5 edits per month) is plotted over time.
Wikipedia is great. But it’s got issues:

- Hard to attract and retain contributors—especially the diverse and marginalized.
- These problems may drive “knowledge gaps.”
- Wikipedia’s quality control system is involved.
An interface for moderating Wikipedia

ORES Flags

- (diff | hist) . . Valletta-Floriana rivalry; 19:31 . . (+38) . . 77.71.194.111 (talk) (→Cultural rivalry)
- (diff | hist) . . Delray Beach station; 19:31 . . (-1,437) . . C16sh (talk | contribs) (→Station layout: use template)
Evaluating the fairness of a sociotechnical system

Screenshot of the front page of Selbst et al. “Fairness and Abstraction in Sociotechnical Systems,” removed for CC-BY-SA.

https://doi.org/10.1145/3287560.3287598
How can we evaluate algorithmic flagging in a sociotechnical system?

1. Understand the system and its normative standards (i.e., fairness).
2. Measure outcomes of the system in terms of the normative standards.
Sources of bias

Moderators use social signals like registration and experience to find and sanction misbehavior.

People displaying such signals might be over-profiled if moderators focus their attention on them but not on others engaged in similar behavior.

Algorithms might be biased against these very same users further exacerbating over-profiling.
ORES is biased against unregistered editors

![Graph showing bias against unregistered editors in various Wikipedias](image_url)
ORES is less biased against editors without user pages.
Decision system fairness metrics: quantify fairness of decisions (in contrast to fairness of predictions)

Demographic parity: Who you are doesn’t affect how you’re treated.

Meta-norms: Define right and wrong ways of sanctioning. Violations of meta-norms may be controversial sanctions.

False positive rate (FPR): How often sanctions violate meta-norms.

Equality of opportunity: Who you are doesn’t affect the FPR.
Research questions

Will flagging increase or decrease demographic parity?
Will flagging increase or decrease the FPR for overprofiled editors?
Will flagging increase or decrease equality of opportunity?
Research questions

Will flagging increase or decrease demographic parity?
Will flagging increase or decrease the FPR?
Will flagging increase or decrease equality of opportunity?
Data: Wikimedia History

ORES Flags

- (diff | hist) . . Valletta-Floriana rivalry; 19:31 . . (+38) . . 77.71.194.111 (talk) (→Cultural rivalry)
- (diff | hist) . . Delray Beach station; 19:31 . . (-1,437) . . C16sh (talk | contribs) (→Station layout: use template)

Red user profile link

Unregistered editor

- Public data of Wikipedia edit history
- Historical prediction scores maintained by Wikimedia (to be released)
- Thresholds (maybe damaging, likely damaging, very likely damaging) reconstructed from old models and configuration files
- Stratified sample with up to 23 different language editions of Wikipedia.
Regression discontinuity design
Regression discontinuity design
Regression discontinuity design
Research design: regression discontinuity

An algorithmic predictor triggers flags when prediction scores an *arbitrary threshold*.

So we infer the causal effects of flagging on moderation by comparing edits right above the threshold with and conditioning on the scores.

![Graph showing the probability of reverted edits against distance from the threshold](image-url)
Results
RQ1: Flagging and over-profiling: Registration status

Distance from threshold

Prob. Reverted

Threshold
- Maybe damaging
- Likely damaging

Unregistered
-0.03 0.00 0.03
0.10
0.15
0.20
0.25

Registered
-0.03 0.00 0.03
0.05
0.10
0.15
0.20
RQ2: Flagging and fair sanctioning: Unregistered editors

Flagging decreases controversial sanctioning for unregistered editors.
RQ1: Flagging and over-profiling: User pages

Distance from threshold

Prob. Damaging

Threshold

- Maybe damaging
- Likely damaging

No user page

User page
We don’t detect a change in controversial sanctioning for editors without user pages.
Conclusion

Perhaps editors without User pages are not over profiled, or the algorithm is biased against them.

While our estimates of the effects of flagging are causal, comparison between types of editors is not.

This is about moderation on Wikipedia, but social and psychological processes might be similar in high stakes settings.

While algorithmic flagging can improve fairness for some classes of user, the general relationship seems complex and contingent.
Evaluating fairness of a sociotechnical system depends on understanding the system.

Decision system fairness metrics should capture emic values of the system.

Regression discontinuity designs provide a non-interventional tool for evaluating flagging systems.

How algorithmic flagging shapes decision system fairness may be difficult to predict.
Thank you!

Thanks to the NSF GRFP #2016220885 and the Wikimedia Foundation. 
nathante@uw.edu

@groceryheist

Link to paper: https://dl.acm.org/doi/abs/10.1145/3449130

My website: https://teblunthuis.cc

Research group’s website: https://communitydata.science