THE SANTA CLARA PRINCIPLES

On Transparency and Accountability in Content Moderation

These principles are meant to serve as a starting point, outlining minimum levels of transparency and accountability that we hope can serve as the basis for a more in-depth dialogue in the future.

On the occasion of the first Content Moderation at Scale conference in Santa Clara, CA on February 2nd, 2018, a small private workshop of organizations, advocates, and academic experts who support the right to free expression online was convened to consider how best to obtain meaningful transparency and accountability around internet platforms' increasingly aggressive moderation of user-generated content.

Now, on the occasion of the second Content Moderation at Scale conference in Washington, DC on May 7th, 2018, we propose **these three principles** as initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights.

ACLU Foundation of Northern California

Center for Democracy & Technology

Electronic Frontier Foundation

New America's Open Technology Institute

Irina Raicu

Markkula Center for Applied Ethics, Santa Clara University

Nicolas Suzor

Queensland University of Technology

Sarah Myers West

USC Annenberg School for Communication and Journalism

Sarah T. Roberts

Department of Information Studies, School of Education & Information Studies, UCLA

1 Numbers

Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.

At a minimum, this information should be broken down along each of these dimensions:

- Total number of discrete posts and accounts flagged.
- Total number of discrete posts removed and accounts suspended.
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by category of rule violated.
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by format of content at issue (e.g., text, audio, image, video, live stream).
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by source of flag (e.g., governments, trusted flaggers, users, different types of automated detection).
- Number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by locations of flaggers and impacted users (where apparent).

This data should be provided in a regular report, ideally quarterly, in an openly licensed, machine-readable format.

2 Notice

Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.

In general, companies should provide detailed guidance to the community about what content is prohibited, including examples of permissible and impermissible content and the guidelines used by reviewers. Companies should also provide an explanation of how automated detection is used across each category of content. When providing a user with notice about why her post has been removed or an account has been suspended, a minimum level of detail for an adequate notice includes:

- URL, content excerpt, and/or other information sufficient to allow identification of the content removed.
- The specific clause of the guidelines that the content was found to violate.
- How the content was detected and removed (flagged by other users, governments, trusted flaggers, automated detection, or external legal or other complaint). The identity of individual flaggers should generally not be revealed, however, content flagged by government should be identified as such, unless prohibited by law.
- Explanation of the process through which the user can appeal the decision.

Notices should be available in a durable form that is accessible even if a user's account is suspended or terminated. Users who flag content should also be presented with a log of content they have reported and the outcomes of moderation processes.

3 Appeal

Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

Minimum standards for a meaningful appeal include:

- Human review by a person or panel of persons that was not involved in the initial decision.
- An opportunity to present additional information that will be considered in the review.
- Notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision.

In the long term, independent external review processes may also be an important component for users to be able to seek redress.

Acknowledgements

We thank Santa Clara University's High Tech Law Institute for organizing the Content Moderation & Removal at Scale conference, as well as Eric Goldman for supporting the convening of the workshop that resulted in this document. That workshop was also made possible thanks to support from the Internet Policy Observatory at the University of Pennsylvania. Suzor is the recipient of an Australian Research Council DECRA Fellowship (project number DE160101542).