THE SANTA CLARA PRINCIPLES

ON TRANSPARENCY & CONTENT MODERATION

1 NUMBERS

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

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.

3 APPEAL

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

To develop these recommendations, partners undertook a thematic analysis of 380 survey responses submitted by users to EFF's onlinecensorship.org who have been adversely affected by the removal of content they pose on social media platforms or by the suspension of their account. These principles build on this wider research process as well as the deliberative sessions at the **All Things in Moderation conference** at UCLA (6-7 December 2017). The research was used to identify information gaps expressed by users about what content is moderated, which rule was breached, and what human and automated processes are responsible for identifying content and making decisions about content moderation.

1 NUMBERS

COMPANIES SHOULD PUBLISH THE NUMBER OF POSTS REMOVED AND ACCOUNTS PERMANENTLY OR TEMPORARILY SUSPENDED DUE TO VIOLATIONS OF THEIR CONTENT GUIDELINES

At 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** (eg. 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** (eg. 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 location of flaggers and impacted users** (where apparent).

This information should be provided in a regular report, ideally quarterly and in an openly licensed and 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 a post has been removed or an account has been suspended, a minimum level of detail 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 governments 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 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 included 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 workshop that resulted in this document. The workshop and research were 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 no. DE160101542).

PARTICIPANTS

- 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 T. Roberts (Department of Information Studies, School of Education & Information Studies, UCLA)
- Sarah Myers West (USC Annenberg School for Communication and Journalism)