GitHub

88 Colin P Kelly Jr Street,
San Francisco, CA 94107
Tel: 415-448-6673 (main)

June 30, 2020

via email SCPconsultation@eff.org

Re: **Review of Santa Clara Principles on Transparency and Accountability in Content Moderation**

GitHub welcomes the opportunity to contribute to the review of the Santa Clara Principles on Transparency and Accountability in Content Moderation.

GitHub is the world's largest software development platform, enabling users and businesses to collaboratively develop open-source and proprietary software projects. GitHub's global community of over 50 million users includes individual developers, startups, small businesses, large companies, NGOs, and governments. GitHub-hosted software projects include applications designed for web or mobile devices, as well as the source code that powers entire businesses. Developers on GitHub work together, sharing code and knowledge, to build the future of software. GitHub hosts content in over 100 million repositories.

We offer this submission from the perspective of a software development collaboration platform hosting user-generated content and whose users include developers building and maintaining user-generated-content platforms.

Our key recommendations include to

- note that a platform's policies that are the basis for violations should be clear
- add partial restrictions to the scope of the principles
- encourage proportionality vs. a one-size-fits-all approach
- consider where user self-moderation may fit into the principles' scope
- clarify the meaning of a few terms (like "flagged" and "appeal").

When thinking about proportionality, considerations should include the size and resources of the company, size of the platform, volume of content moderation issues the platform encounters, and the platform's risk profile.

1. Please indicate any specific recommendations or components of the category that should be revisited or expanded.

- <u>publish the numbers</u> of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines

- We recommend clarifying the meaning of "flagged," for example, flagged for review or flagged as removed.

- We recommend clarifying what constitutes a "discrete post," as posts can be represented in many ways given the various ways users may interact with a platform.

- We recommend revising the "Numbers" principle to incorporate a degree of proportionality in its various components. Numbers can be a useful way to understand aspects of how platforms moderate content. However, not all platforms are well equipped to track and report with granularity, and the story of these numbers is more telling for some platforms than others. For example, the current category components make sense for large platforms that have a high volume of commonly reported-on content moderation issues. The level of transparency sought through numbers in this principle's components should be proportionate to the platform's level of content moderation and risk profile, so that platforms that process a high volume of takedowns of a particular kind of violation would track statistics on those—without necessarily having to track and report on all.

- As we explain in question 8, content moderation often (and should) entail actions beyond a binary decision to remove content or suspend an account. Tracking partial-restriction actions such as geoblocking and de-ranking content can also be useful in showing a platform's content moderation practices.

- For larger platforms operating in many locales, a regional breakdown of content moderation actions is often useful.

- <u>provide notice to each user</u> whose content is taken down or account is suspended about the reason for the removal or suspension

- Relevant to notice, we recommend adding that the reason given for an action should only be based on a published policy.

- We also recommend considering whether to recommend advanced notice, for example, where the affected user might be able to take an action to address the issue that would obviate the need to remove content or to restrict access to an account.

- As for providing a detailed response (the specific clause violated and information about how the platform identified the content) as part of the notice, we recommend allowing for exceptions in certain cases, such as spam and phishing (as recognized in EFF's 2018 Who Has Your Back Report), as well as accounts that appear to be created solely with a purpose that violates a platform's terms (like harassing other users). Providing a detailed response in these situations can be counterproductive, particularly because these users may not otherwise try to use their account again. To

be clear, we recommend these exceptions specifically with respect to providing a detailed response, not to providing notice generally. Users should be notified, with a way to contact a platform, even in those exceptions.

- Regarding "Users who flag content should also be presented with a log of content they have reported and the outcomes of moderation processes," this makes sense for large platforms. However, the ability to do this kind of tracking and follow-up with users can be especially burdensome, particularly in terms of time it takes away from reviewing other potential violations, and often requires product decisions and engineering resources to prioritize and support trust and safety. This is especially true for platforms with complex structures for hosting content, as opposed to those with one simple structure such as a timeline or a newsfeed.

- provide a meaningful opportunity for timely appeal of any content removal or account suspension

- We recommend defining "appeal" and allowing for exceptions in certain cases, for example, for spam, phishing, and malware.

- We suggest a clarification that "meaningful opportunity" would apply where there is new information, but not where someone responds demanding more and more people review without providing any no new information for them to consider.

2. Do you think the Santa Clara Principles should be expanded or amended to include specific recommendations for transparency around the use of automated tools and decision-making (including, for example, the context in which such tools are used, and the extent to which decisions are made with or without a human in the loop), in any of the following areas:

- **Content moderation (the use of artificial intelligence to review content and accounts and determine whether to remove the content or accounts; processes used to conduct reviews when content is flagged by users or others)**
- **Content ranking and downranking (the use of artificial intelligence to promote certain content over others such as in search result rankings, and to downrank certain content such as misinformation or clickbait)**
- **Ad targeting and delivery (the use of artificial intelligence to segment and target specific groups of users and deliver ads to them)**
- **Content recommendations and auto-complete (the use of artificial intelligence to recommend content such as videos, posts, and keywords to users based on their user profiles and past behavior)**

We recommend that the principles address transparency and accountability around automated content moderation given the growing and permanent importance of automation in content moderation. We encourage care in doing so in order to (1) keep the principles focused on content *moderation* rather than expanding in scope to cover any and all decisions about content display and distribution and (2) to the extent scope is increased, consider feasibility for developers and platforms in what the principles recommend. For example, informing users of the types and nature of automated content decisions being made by a platform might be a baseline. Further measures such as developers instrumenting all code involved in the display and distribution of content for the purposes of transparency reporting,

or open sourcing code and models that are used to make decisions about content, may be desirable for some platforms, but would be infeasible for many others.

3. Do you feel that the current Santa Clara Principles provide the correct framework for or could be applied to intermediate restrictions (such as age-gating, adding warnings to content, and adding qualifying information to content). If not, should we seek to include these categories in a revision of the principles or would a separate set of principles to cover these issues be better?

We recommend incorporating intermediate restrictions into the existing fundamental content moderation considerations in the Santa Clara Principles because in many cases, a more nuanced action is appropriate and proportionate to the problem raised by the content. From a logistical perspective, given how much momentum it takes to draft, publish, and gather support for principles, as well as to track their implementation and keep them current, it probably makes sense to revise these principles to include them on that basis too.

Examples of intermediate restrictions based on our platform include

- more granular content removal (for example by giving a user the opportunity to fix by removing specific content rather than removing or disable an entire page or project)
- suspend a user without disabling their content
- disable a repository vs. an account
- geoblock vs. globally block
- downgrade visibility or discoverability, such as by blocking from non-logged-in users or de-indexing from search engines
- set temporary interaction limits.

4. How have you used the Santa Clara Principles as an advocacy tool or resource in the past? In what ways? If you are comfortable with sharing, please include links to any resources or examples you may have.

We noted our public support of the principles in GitHub's 2018 Transparency Report and described it in our blog post about EFF's 2019 Who Has Your Back Report, which gave a star to companies that publicly support the principles.

5. How can the Santa Clara Principles be more useful in your advocacy around these issues going forward?

Greater awareness that companies are incorporating them into their operations could help counter arguments that tech companies don't care about users and that they treat moderation in an unfair, biased, or capricious manner.

6. Do you think that the Santa Clara Principles should apply to the moderation of advertisements, in addition to the moderation of unpaid user-generated content? If so, do you think that all or only some of them should apply?

Numbers, notice, and appeals are key aspects of fairness and are often relevant beyond the context of unpaid user-generated content.

7. Is there any part of the Santa Clara Principles which you find unclear or hard to understand?

As we noted in question 1, we recommend defining the meaning of “flagged” and “appeal” and clarifying the meaning of “posts” in “discrete posts.”

We would be interested to know under what circumstances advanced notice might be encouraged for content removals. (See also question 1 for circumstances to consider.)

We would also encourage more information and guidance about how smaller enterprises can get started with the principles short of implementing all of the principles to a tee.

8. Are there any specific risks to human rights which the Santa Clara Principles could better help mitigate by encouraging companies to provide specific additional types of data? (For example, is there a particular type of malicious flagging campaign which would not be visible in the data currently called for by the SCPs, but would be visible were the data to include an additional column.)

Similar to question 3, above, the Santa Clara Principles currently focus on flagging, suspension, and removal of content. Reporting that data is helpful but would not show human rights risks reflected in partial-restriction action such as geoblocking, de-ranking, or other moderation action short of flagging, suspension, or removal.

A column or two to reflect non-removal moderation action would help show potential malicious campaigns with value for human rights defenders. For example, a country could target certain kinds of speech that a company might geoblock but may not remove globally. This may not show in transparency expectations under the current Santa Clara Principles, but it could still violate speech rights in a given country or region.

While not a “specific type of data,” another good practice we recommend is for platforms to publicly post notices they act on—notifying reporting users or entities that they will do so and redacting appropriately—when taking action against a user on the basis of a law. For example, GitHub publishes notices we process from governments seeking content removal. As another example, GitHub publishes takedown notices and counter notices we process under the Digital Millennium Copyright Act (DMCA). Publishing the content of the notices helps to deter overzealous reporting and promotes transparency by enabling the public to know the basis for a platform’s takedown decision. While we recommend this practice for takedowns based on a law, we caution against that practice for abuse reports given the complexity and sensitivity of reports, as well as the need to protect users and reporting users.

Similarly, when disabling content, where possible, we recommend that platforms note the reason (for example “DMCA takedown”) as the error message rather than a generic 404 error.

9. Are there any regional, national, or cultural considerations that are not currently reflected in the Santa Clara Principles, but should be?

Given the global nature of the internet and of many userbases, it could be useful to recommend that moderators have language and/or cultural competency to moderate content of users from different regions, countries, or cultures than their own, as relevant to the nature of content on the platform and degree of risk to users. This is particularly true for larger

companies where a large percentage of a population uses their platform as a means of communication on social issues. (While a good idea across the board, it can be difficult for small and medium enterprises to achieve given the proportionately small size of their teams.) It could also be useful for a platform to indicate they have limited competency where they do, for example, noting that content moderators provide support in English only if that's the case.

10. Are there considerations for small and medium enterprises that are not currently reflected in the Santa Clara Principles, but should be?

Some of the benchmarks that involve tooling to accomplish can be difficult for small and medium enterprises to build or buy, particularly if the risk of a problem with content moderation doesn't justify it. The principles should not be a one-size-fits all prescription, but rather aspirational guidelines that small or medium-sized platforms can adapt to meet their needs.

Perhaps the principles could add a general statement regarding proportionality of meeting them based upon the size and resources of the company, size of the platform, and volume of content moderation issues the platform encounters, as well as its risk profile. As we noted in question 1, there are likely opportunities to consider proportionality in some of the principles, particularly in the "Numbers" principle.

On a related note, if the Santa Clara Principles were to address timelines for removal, this is another area that presents a barrier to competition and success for small and medium enterprises because they often lack the resources to comply, so either would often fail to meet strict deadlines, or would err on the side of removing content even if not a violation. Where policymakers seek fast action, a way to level the playing field and lessen human rights risks to free expression is to only call for faster action where notices originate from a government entity, and not from any reporting user.

11. What recommendations do you have to ensure that the Santa Clara Principles remain viable, feasible, and relevant in the long term?

- Periodic reassessment and amendment, like this one
- Continuing application to new content moderation developments that have gone well and that haven't
- Continuing assessment by diverse stakeholders
- An auditing process for platforms

12. Who would you recommend to take part in further consultation about the Santa Clara Principles? If possible, please share their names and email addresses.

13. If the Santa Clara Principles were to call for a disclosure about the training or cultural background of the content moderators employed by a platform, what would you want the platforms to say in that disclosure? (For example: Disclosing what percentage of the moderators had passed a language test for the language(s) they were moderating or disclosing that all moderators had gone through a specific type of training.)

If the Santa Clara Principles were to call for such disclosures, they should correspond to the nature of the content moderated on the platform and its userbase, and should only be in the aggregate to protect the safety and privacy of content moderators.

14. Do you have any additional suggestions?

To understand the full scope of content moderation on a given platform, in addition to tracking a platform's staff moderation, it can help to track how users self-moderate their own spaces. This may be hard to standardize across platforms, but for a platform like GitHub, for example, we could consider tracking repositories that have codes of conduct—as an indicator of setting clear policies or terms that are the basis for enforcement actions—and providing visibility into moderation actions such as users blocking other users. Any categories related to self-moderation should reflect that a platform may not be able to track self-moderation actions as easily as its own staff moderation actions.

15. Have current events like COVID-19 increased your awareness of specific transparency and accountability needs, or of shortcomings of the Santa Clara Principles?

A global shift toward heavier reliance on online social platforms, remote meeting technology, and other hosted solutions to enable safe living—both professional and personal—during a pandemic absolutely increases the importance of robust transparency and accountability guidelines for online platforms. It also highlights the weight of content moderation decisions that platforms face when reported content could be a government's communication to its residents about important health information or disinformation. These kinds of best practice guidelines and frameworks are key to assuring continued flexible, pragmatic self-governance, and avoiding heavy-handed, disproportionate, or restrictive compliance burdens.

***

GitHub thanks you again for the opportunity to contribute to the review of the Santa Clara Principles and to offer recommendations from the perspective of a software development collaboration platform hosting user-generated content and whose users include builders and maintainers of user-generated-content platforms. We look forward to the outcomes of this process.