Santa Clara Open Consultation Report

This report was written in collaboration by: Access Now, ACLU Foundation of Northern California, ACLU Foundation of Southern California, ARTICLE 19, Brennan Center for Justice, Center for Democracy & Technology, Electronic Frontier Foundation, Global Partners Digital, InternetLab, National Coalition Against Censorship, New America's Open Technology Institute, Ranking Digital Rights, Red en Defensa de los Derechos Digitales, and WITNESS

A publication of the Electronic Frontier Foundation, 2021.
"Santa Clara Principles Open Consultation Report" is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

View this report online: https://santaclaraprinciples.org/open-consultation/

Santa Clara Principles Open Consultation Report

DECEMBER 8, 2021

Executive Summary

Throughout the past few years, as tech companies have taken on an increasingly complicated role in policing the world's speech, the demand for greater due process and transparency in content moderation has grown. At the same time, and particularly during the pandemic, gains made in recent years on transparency and accountability measures have in some instances regressed.

The call for transparency from social media companies dates to more than a decade ago when, following Yahoo!'s handing over of user data to the Chinese government—an act that resulted in the imprisonment of local dissidents Wang Xiaoning and Shi Tao—collective outrage from digital rights activists in organizations led to the creation of the Global Network Initiative (GNI), a multi-stakeholder organization that seeks to hold tech companies accountable to a set of principles, one of which is public transparency.

This effort, along with others, resulted in the publication of the first transparency reports by major companies such as Google. It is because of these transparency reports that we know how many pieces of user data companies turn over to various governments, or how many occurrences of government censorship have been imposed on platforms. Such transparency enabled users to make more informed decisions about which products to use and avoid, and empowered advocacy groups to insist that companies follow established legal processes when complying with such government demands.

Over time, however, it became clear that transparency around how companies respond to governments was insufficient; civil society began to demand that companies provide information about how they enforce their own policies and practices. The formation of Ranking Digital Rights—an organization which works to promote freedom of expression and privacy on the internet by creating global standards and incentives for companies to respect and protect users' rights—and the launch of their Corporate Accountability Index in 2015 pushed the field forward, as did other efforts such as the Electronic Frontier Foundation's Who Has Your Back?, a project that ranks companies based on their adherence to an annually updated set of transparency standards.

In 2018, a small convening of academics, lawyers, and activists culminated in the creation of the Santa Clara Principles on Transparency and Accountability in Content Moderation, a clear set of baseline principles designed to obtain meaningful transparency around internet platforms' increasingly aggressive moderation of user-generated content. The principles comprise three categories: Numbers, Notice, and Appeal. Each category is then broken down into more detail.

The "Numbers" category asks companies to publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines. "Notice" indicates that companies should provide ample notice to users when their content is removed or their account suspended, with sufficient detail about

the policy(ies) violated and information on how to appeal. Finally, "Appeal" asks that companies provide a meaningful opportunity for timely appeal, with a set of minimum standards that emphasizes human review.

In 2018, in the wake of several high-profile censorship events by Facebook, a coalition that included several of the original authors penned an open letter to Mark Zuckerberg demanding that the company institute an appeals process across all of its policy categories. That effort succeeded in part, opening a broader dialogue with Facebook and resulting in the expansion of their appeals processes to cover most areas of policy. In 2019, the Electronic Frontier Foundation's "Who Has Your Back?" project was successful in obtaining endorsements of the principles by twelve major companies, including Apple, Facebook, Twitter, Reddit, and YouTube. Unfortunately, however, only one company—Reddit—has implemented the principles in full and the three largest platforms—Facebook, Youtube, and Twitter—were significantly lacking in their implementation.

At the same time—and particularly as many companies have taken an increasingly aggressive and often automated approach to content moderation, particularly in the Global South—the original group of authors heard feedback from many of our allies working on content moderation and digital rights around the world that the original principles missed certain key issues and could benefit from a review. As such, we undertook a lengthy deliberative process that involved more than fifteen organizations and an open call for comments with the goal of eventually expanding the Santa Clara Principles.

The pandemic brought with it a number of new complexities. In March 2020, as workplaces shut down and many countries went into lockdown, content moderators were by and large sent home and replaced with automated processes, many of which are buggy at best and deeply insufficient at worst. Many of the organizations working in our field have reported a great increase in the number of complaints they receive from users seeking redress and assistance in regaining their accounts. In April 2020, a global group of more than 40 organizations addressed the companies, urging them to ensure that removed content would be preserved, decisions made transparent, and that information would, in the future, be given to researchers and journalists.

At the same time, we are seeing an increase in demands for censorship, both by the public and governments, the latter of which in particular has resulted in what appears to be widespread silencing of dissent and opposition movements in various locales, including India, Palestine, the United States, and Colombia.

The pandemic has also directly affected our own work. Our plans to organize a convening alongside RightsCon2020 where we could receive in-person feedback from digital rights activists were dashed when travel was grounded. Furthermore, many organizations have struggled with capacity issues over the past year. Nevertheless, we received nearly thirty in-depth submissions from groups and individuals in roughly eighteen countries, from Brazil to Kenya to Canada, including several real-time group consultations conducted by partners in Africa, Latin America, India, and North America.

These submissions were thoughtful, nuanced, and reflect a wide range of views on how transparency and accountability efforts by companies can be expanded to benefit a diverse range of users.

This report is reflective of those submissions and observes the following trends:

- A desire amongst respondents for greater due process and respect for human rights frameworks throughout the content moderation process
- A demand for cultural competency of content moderators throughout the content moderation process, with a specific need for understanding of local cultural contexts
- A need for transparency about the use of automated systems used in both content moderation and in recommendation of content
- A demand for human oversight of automated process and human review of content appeals
- A clear desire for transparency around government involvement in the moderation of content and users
- A desire for robust requirements, as contained within the principles, to be adopted as an online service provider matures.
- A desire for transparency around individual content moderation decisions and data that can assist in identifying systemic trends and effects

The Original Santa Clara Principles

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.

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 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 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.

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.

Results of the 2020-21 Open Global Consultation

A. Overarching and Cross-Cutting Comments and Recommendations

The open consultation revealed several over-arching and cross-cutting comments and recommendations.

1. Broadening the definition of "Content Moderation" beyond takedowns and account suspensions

The initial version of the Santa Clara Principles focused on a limited set of moderation decisions: the removal of posts and the suspension of accounts, both temporary and permanent.

But the taxonomy of content moderation is full of numerous other types of "actioning," the term used by platforms to describe a range of enforcement actions including removal, algorithmic downranking, and more. Commenters were asked whether the Principles should be extended to a range of other decisions.

The overwhelming majority of respondents supported extending the Principles to other types of moderation decisions, both intermediate moderation actions such as downranking, as well as AI-based content recommendation and auto-complete. Some respondents reasoned that users should have greater control over what is shown to them; others believed that if certain content is recommended, other pieces of content are less likely to be seen, raising the risk of discrimination as to who sees what content. Still others argued, however, that content promotion/downranking was sufficiently different from content removal to require a different approach, and even those who favored including such actions did not have specific recommendations for what companies should be required to do. Several respondents highlighted the importance of defining the term "intermediate restrictions" so, at a minimum, it is clear to what actions the Santa Clara Principles apply.

Recommendations

- ★ As revised, the new principles apply broadly to negative actions taken by companies in response to either the content of users' speech or the users' identity. The revised principles apply to all "actioning" of user content by companies, defined as any form of enforcement action taken by a company with respect to a user's content or account due to non-compliance with their rules and policies, including (but not limited to) the removal of content, algorithmic downranking of content, and the suspension (whether temporary or permanent) of accounts.

- ★ We also explored extending the Santa Clara Principles to autocomplete and recommendations systems. However, we decided not to include auto-complete and recommendations systems within these Principles since to do so seemed unworkable given how frequent such actions are and the varying contexts in which they occur.

2. Due Process Throughout The Content Decision-Making System

While implicit, the first version of the Santa Clara Principles did not specify steps that companies engaged in content moderation should take before and in making the initial moderation decision, as opposed to the appeal of that decision. Many of the comments received during the consultation were directed at that first level of moderation, with commenters wanting greater assurances of transparency and due process throughout the entire process of moderation. Indeed, much of the desire for transparency around the first-level decision was founded in the lack of trust that those decisions were being made fairly, consistently, and with respect for human rights. While transparency is important to evaluating human rights values, it is not itself a substitute for them.

Several comments asked for more detail and clarity around a platform's content policies: although most companies publish their content policies online, these rules are

not always comprehensible. Some companies do not have clearly defined and named policies and categories, which makes it difficult to understand them. In addition, policies often make use of vague and broadly interpreted terms, such as "offensive," raising concerns around overbroad removals. Further, companies do not clearly outline how they implement and enforce their content policies (including which tools they use to moderate content), making it difficult to understand the full spectrum of their content moderation efforts and hold them accountable. Another issue raised relates to remedy measures that platforms should take when reversing content moderation decisions following a successful appeal (e.g. by providing greater content visibility and reviewing their policies).

Commenters also seek greater assurances that both human and automated moderators are sufficiently trained and have sufficient expertise. Commenters wanted to know whether human moderators are employed or outsourced or contractors, how they are supervised, how their performance is assessed so that quality decision-making is optimized, and information about their working conditions.

Specific suggestions included the following:

- "When humans are involved, it is important that we know about their role in the content-moderation process (are they supervising or being supervised?), their location/background and any training process they underwent to prepare them for their roles. Also, to monitor content-curation, it is important to know whether these monitoring teams are internal or outsourced to other companies (which could potentially create a monopoly of content-moderating actors). . . . The bias present in human moderators is bound to affect the data set being used to train the AI moderator too. This worry makes it urgent to know who's doing the moderating and to figure out ways to make the data more representative. We would need to know: the percentage of people undergoing certain training, their role in the content moderation process (Are they supervisors? Are they involved in the appeal process too?); their professional experience (i.e., Are they lawyers?); their location and language; and whether these monitoring teams are internal or outsourced to other companies. It would also be useful to know what are the policies for their protection (e.g. physically, emotionally, mentally, etc.), what are the incentives that are offered to them, how is their performance measured, and other workplace conditions as this will shed light on to how the human content moderators are making their decisions." - *Montreal AI Ethics Institute* (Canada)

- "The following information should be disclosed: All the specific training sessions that the moderators had gone through, along with the time spent on each course; How the effectiveness of content moderation is improving or decreasing with each cycle of training sessions; The evaluation reports filled in by the content moderators concerning the specific training; Lists showing moderators who are employed by certain companies, e.g medical companies, have a certain amount of training/qualifications in the specific field." - *Lawyers' Hub* (Kenya)

- The training provided by the platforms also represents a relevant aspect of the performance of the moderators, which can be subject to additional transparency measures to be adopted by the platforms. An updated version of the Santa Clara Principles could ask platforms to disclose more information about this training, as well as the internal guidelines for moderation that moderators should follow." - *InternetLab* (Brazil)

- "We suggest that principles for content reviewer transparency should focus on processes for hiring and training reviewers. By contrast, a metrics-based approach may, ultimately, provide less valuable information. For instance, assuming that platforms have a minimum standard of language proficiency for content moderation teams, it would be less informative to know that 100% of individuals moderating content in a given language were deemed proficient, and more informative to understand the processes by which reviewers are hired, assessed for language competence, and trained." - *Facebook/Meta (US)*

- "Alongside disclosure regarding the nature of training given to the human moderators and their internal moderation norms, we also recommend that the Principles recognize certain fundamental ethical guidelines with relation to their human moderators that companies must adopt. This might include providing information of the third-party firms to which the company outsources its moderation and assurances of sufficient number of counsellors for the moderators." - *Centre for Internet and Society,* (India)

Recommendations

★ As revised, the new principles include a foundational principle of Human Rights and Due Process that emphasizes the need for due process and respect for human rights throughout the entire content moderation process and an Integrity and Explainability Principle that requires that all content moderation processes have integrity and be administered fairly.

3. Cultural Competence Throughout the System

Perhaps the most consistently identified need among all of the comments received was a need for greater cultural competence—knowledge and understanding of local language, culture and contexts—throughout the content moderation system, with the commenters correctly identifying it as an essential component of due process.

Cultural competence generally requires that those making first level moderation and appeal decisions understand the language, culture, and social context of the posts they are moderating. Commenters thus seek both guidelines for companies and greater transparency around the demographic background of human moderators, their geographical distribution, and the language proficiency of moderators to ensure that such competency is in place.

A few commenters recognized that the collection and disclosure of personal data about moderators raises considerable privacy concerns.

Specific comments included the following:

- "The cultural and social context of content moderators may have implications for the way content policies are interpreted and implemented. Therefore, an updated version of the Santa Clara Principles could recommend the disclosure (and frequent updating) of quantitative data on the community of platform moderators in order to attest that the diversity of its users is also reflected there. Thus, platforms could disclose how many people they hire as content moderators, breaking this number down by jurisdictions in which they operate and by the language they operate in, as well as diversity indications (nationalities, declared gender and ethnicity, for instance). - *InternetLab* (Brazil)

- More than knowing the background of moderators, in order to understand a little about how content review is distributed, it is important to know how many people are reviewing the content in a country to know what moderator resources are available in a given region.. - *Latin America virtual consultation*

- "If Arabic content is taken down, Arabic responses should be issued to users. Companies need to utilize the language of the user to respond." - *7amleh* (Palestine)

- "For how long have the moderators responsible for a certain language actually lived in the area?" - *Taiwan Association for Human Rights* (Taiwan)

- "Given the global nature of the internet and of many user bases, 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.... 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." - *GitHub* (US)

- We need greater transparency and accountability to ensure that moderators have the relevant cultural contexts and expertise in issues such as racism, civil rights,etc. - *Americas virtual consultation*

- "This needs to be revisited with regards to language - what about content in languages other than English? What kind of standards should apply to non-English content, and how can these standards differ according to different languages? How can companies ensure linguistic diversity in content moderation (by having linguistically diverse teams, diversifying data sets, etc.)? .... It should also include information on the composition of the panel, the qualification of the person(s), what communities and what geographical locations are represented in the panel, etc." - *Point of View* (India)

- "More transparency around moderators would be useful, and moderators should come from diverse backgrounds. But the Principles should not focus on the moderators. Moderators make decisions based on the guidelines provided by each company and on content flagged via automation. The target of the Principles would be to have guidelines that are fair to all." - *Association for Progressive Communications* (International)

Recommendations

★ As revised, the new principles establish understanding of local cultures and contexts as a foundational principle that must be considered and aimed for throughout the operational principles. Specific steps are also included within each operational principle.

4. Special Consideration for the Use of Automation in Content Moderation

Internet platforms increasingly rely on automated tools powered by artificial intelligence and machine learning to enhance and manage the scale of their content moderation processes. Companies use algorithmic curation mechanisms, such as downranking and recommending, to moderate content that violates or partially violates their rules. Companies use content filters to scan content before it is published to the service—in some cases preventing the upload of content that violates certain rules.

Although these tools can make it easier for companies to moderate and curate content at great scale, they have also resulted in numerous concerning outcomes, including overbroad takedowns, deamplification of content, and other moderation measures. Automated moderation tools are generally limited in that they are unable to assess context or make subjective decisions. This has had an outsized impact on certain communities; for instance, the use of automation in moderating violent extremist content has arguably resulted in the widespread erasure of evidence of human rights violations in countries such as Syria. As a result, these tools cannot make accurate decisions when flagging or moderating content that doesn't have clearly defined parameters or that requires nuance and specific cultural or linguistic understanding.

We asked stakeholders their opinions on expanding the Santa Clara Principles to specifically include guidelines related to the use of automated tools and AI in detecting and removing posts, for the purposes of ranking content, and for content recommendation and auto-complete.

Commenters strongly supported amending the Santa Clara Principles to account for the ever-increasing use of automated systems in the moderation process, as well as for content ranking.

- "Creating at least some principles specific to human moderators and others specific to AI moderation could also contribute to the SCP remaining valuable and applicable to a large range of companies and their methods of moderation.... For any content that is automatically moderated without a human in the loop, there should be clear documentation of the criteria against which the decision was made and the confidence levels of the system should be recorded for audit purposes." - *Montreal AI Ethics Institute* (Canada)

- "Considering the limitations of these technologies, an updated version of the Principles could recommend the development of metrics to assess the accuracy of different artificial intelligence tools in identifying problematic content and publish them so that the public can measure the risks associated with the use of these technologies and monitor their improvement. These metrics could be developed, for example, by submitting the same samples of content to human moderators and to artificial intelligence tools and comparing the results of both performances." - *Internet Lab* (Brazil)

- "Principles should be expanded to include content removed by automatic flagging, the error rates encountered by the tools, and the rate at which wrongly taken down content is being reinstated. There should be a clearer disclosure of the kind of automated tools they use. Further, collaborative efforts to the effect of using automated tools in content moderation must be done with sufficient consideration to the basic principles of transparency and accountability." - *Centre for Internet and Society* (India)

- "Automated tools should only be used in the flagging process and content removal should be performed only by human beings where possible. Both humans and algorithms are biased, but humans have additional nuance that may lead to lower percentages of discrimination." - *Lawyers' Hub* (Kenya)

One commenter proposed a new principle specifically aimed at the use of automated decision-making tools.

- "The Explanation principle, as a standard to promote algorithmic transparency, may equally fill gaps in other Santa Clara Principles. Its inclusion will thus strengthen systemically the whole structure, bringing higher practicality to it. With regards to the Numbers principle, for instance, the application of algorithmic transparency to the disclosure of aggregated data related to content

> takedown and flagging in platforms' reports will most certainly turn these companies more accountable. Regarding the Appeal principle, algorithmic transparency equally deepens users' capacity to comprehend the proper dimensions of how one's data is processed and, therefore, strengthens one's capacity to ponder whether or not it's necessary to contest specific decisions related to content flagging or takedown. Lastly, algorithmic transparency prevents opacities where companies should inform its users about what content infringes its usage policies, evidently affronting the Reports principle. Higher transparency about the elucidation of how the users' posts infringe a platform's terms of use and more information of which technique was used to spot such disconformity will allow for a better implementation of this principle." - *Laboratório de Políticas Públicas e Internet* (Brazil)*]*

Although normative guidelines regarding whether automated systems should ever be used without any human moderation are appropriate, most of these suggestions received focused on transparency—companies should disclose when they use automated systems and how accurate those systems have proven to be. Specific suggestions regarding transparency reporting will be addressed below in the section on the Numbers Principle.

Recommendations

- ★ As revised, the new principles have specific provisions regarding the use of automated systems and transparency about that use, as set forth below.
- ★ As revised, the new principles require that companies have high confidence in all systems used, including human, review, automated systems, and all combinations of them. Users should have the ability to seek human review of many moderation actions.

5. Government Engagement with Content Moderation Processes

Government actors typically engage with the content moderation process in several ways: by insisting that community standards and terms of service reflect local law; by submitting legal requests to restrict content based on local law; by flagging legal content that violates the platform's rules through mechanisms designed for end-users; or by utilizing extra-legal backchannels such as Internet Referral Units (IRUs). While companies have long included details about the government takedown request in their transparency reporting, the other practices are often conducted without any transparency.

Respondents strongly agreed that governmental involvement in a platform's content moderation processes raises special human rights concerns, and that those concerns should be specifically addressed in any revision to the Santa Clara Principles. Respondents suggested that platforms provide more transparency generally on

measures that they take in response to government demands, including agreements that may be in place, more information about government pressure or requests, and whether a content moderation action was required by a state regulation. Stakeholders also noted, however, the potential benefits of cooperation between companies and government experts in areas such as election administration and public health.

- "If companies cooperate with governments during a specific event, e.g., covid-19 or election campaign or social movement, companies should publish the cooperation and the number of posts, groups and accounts that were influenced." - *Taiwan Association for Human Rights* (Taiwan)

- Individual participants in the consultations in Africa and India noted the importance of knowing when takedown requests originated with government actors, noting the prevalence of government-affiliated troll armies in theri regions. In the Americas consultation, a participant sought greater transparency about government requests that targeted the speech of journalists and activists. - *Africa, India, and Americas consultations*

- "While the Principles provide a robust framework for content moderation practices carried out by the companies itself, we believe that the framework could be expanded significantly to include more detailed metrics on government requests for content takedown.... For government requests, this information should include the number of takedown requests received, the number of requests granted (and the nature of compliance - including full, partial or none), the number of items identified in these requests for takedown, and the branch of the government that the request originated from (either from an executive agency or court-sanctioned). Information regarding account restrictions, with similar levels of granularity, must also form a part of this vertical. These numbers must be backed with further details on the reasons ascertained by the government for demanding takedowns, i.e. the broad category under which content was flagged." - *Centre for Internet & Society* (India)

- "Regarding government requests, companies do report on the number of pieces of content restricted per country, but they do not consistently disclose the overall number of requests received globally. Having this additional information would make it possible to understand whether requests are on the rise, and the extent to which the company is complying with, or pushing back against, government requests – this would be critical for monitoring and accountability." - *Association for Progressive Communications* (International)

- "Content flagged by government should be identified as such, unless prohibited by law.... It would also be useful to break down which government institutions or public servers, for example: judiciary power, public institutions, intelligence, police and/or military, etc." *Fundación Acceso* (Costa Rica/Central America)

Recommendations

★ As revised, the new principles include a foundational principle addressing state involvement in content moderation that addresses this particular concern.

★ As revised, the new principles require specific reporting of numbers regarding governmental involvement, and additional notice to users, as set forth in the Notice section below.

★ As revised, the new principles include directions to states, namely that they regularly report their involvement in content moderation decisions, including specific referrals or requests for actioning of content, not exploit or manipulate companies' content moderation systems to censor dissenters, political opponents, social movements, or any person, and affirming states' obligations to respect freedom of expression. States must recognize and minimize their roles in obstructing transparency by the companies.

6. A Scaled Set of Principles?

As we sorted through the comments that largely suggested numerous additional reporting and process requirements, we were mindful of a familiar tension: that by increasing content moderation reporting, process, and appeal requirements we risk making them extremely difficult to meet, except for, perhaps, by the current resource-rich, market-dominant intermediaries that uniquely have the resources to fulfill such requirements. Such requirements may then discourage innovation and competition, thus entrenching the existing dominant companies. On the other hand, we recognized that even newer and smaller intermediaries should incorporate human-rights-by-design as they roll out new services. Any new set of principles must be sensitive to avoid generating standards that are impossible for small and medium-sized enterprises (SMEs) to meet compared to large companies. Yet newer and smaller services must understand and plan for compliance with the standards as they scale up, and not wait until they control the speech of millions of users. This concern runs through almost all of the suggested revisions.

Commenters frequently noted the challenge of scaling the Santa Clara Principles so that they were relevant to companies of varying sizes and resources. One participant in the Americas consultation explained that in addition to the issues of scale, strict reporting requirements often prevent companies from publishing data that outlines insights that are unique and most relevant to their type of service.

- Small and medium enterprises "may not have the infrastructure or resources to adhere to the SCP, and appropriately collect and disclose the required information. While this is not an insurmountable barrier, the SCP should at least acknowledge the reality of SMEs and provide some guidance as to what measures they can take to abide by the principles." - *Montreal AI Ethics Institute* (Canada)

- "Our understanding at this current juncture is that not enough data exists around the economic costs of setting up the transparency and accountability structures. Accordingly, at the end of this Consultation period, should the Principles be expanded to include more intermediate restrictions and develop accountability structures around algorithmic use, we recommend that a separate consultation be held with small and medium enterprises to identify a) whether or not there would be any economic costs of adoption and how best the Principles can accommodate them, and b) what are the basic minimum guidelines that these enterprises would be able to adopt as a starting point." - *Centre for Internet & Society* (India)

- "It is good to distinguish responsibilities that should apply to all responsible platforms from those that are only feasible or necessary when applied to dominant platforms." - *Public Knowledge* (US)

- "it is necessary to consider that the size of the platform may influence its capacity to meet the recommendations introduced by the Santa Clara Principles. An updated version of the Principles should make reference to this issue, recommending, for example, longer deadlines for the publication of transparency reports by smaller enterprises." - *InternetLab* (Brazil)

- "Particularly given the likely expanded scope and granularity of the Principles, it may be necessary to provide a simple framework for iterative implementation and the demonstration of progress toward meeting the principles by small and growing enterprises. For example, a basic prioritization of requirements or a simple maturity model that enterprises can self-certify against might prove useful." - *PEN America* (US)

The original Santa Clara Principles set minimum standards and we are mindful that raising them might create unscalable burdens for smaller and new companies that do not have the staff and other resources to provide numbers, notice, and appeals procedures at the same level as the currently dominant platforms. We are also mindful that if we are to undertake scaling, the metrics themselves are problematic. What are the correct measures—Number of users? Capitalization? Geographic reach of service? Extent of moderation? Maturity of service?

We were also mindful that a fixed set of specific requirements may prove overly rigid given how quickly content moderation practices change and the ecosystem evolves.

We considered several ways to approach this problem, such as: tiering the principles such that new requirements were triggered when certain benchmarks were met; keeping the principles as baseline guidelines and providing guidance on how those baselines should be scaled up; and considering whether proportionality could be built into the principles in a way that would provide adequate guidance to companies and adequate protections to users.

Ultimately, we believe that the Santa Clara Principles must establish standards against which to evaluate a company's practices—not minimum standards that everyone must meet, but a mean of sufficient practices. Some companies will and should be able to do more; some will be unable to meet them all. What any one company should do will depend on many factors -- age, user base, capitalization, focus, and more—and will evolve over time.

Recommendations

- ★ As revised, the new principles establish standards that emphasize the purpose of each principle, which must be front of mind regardless of scale. That is, the principles should emphasize the goals of numbers, notice, and appeals, and how each specific practice furthers those goals. The revised principles also include more robust requirements to be adopted as an online service provider matures. The revised principles are also supplemented with a toolkit that provides more specific guidance for companies to consider as they plan for their growth and maturation to ensure that such measures are adopted as companies mature, not after they do so.

The following sections report the range of comments received, and largely do not include an evaluation of the merits or an endorsement of any particular comment.

B. Numbers Principle

1. Summary of Comments

During the consultation process, the Santa Clara Principles coalition solicited feedback on a range of questions which sought to understand if, and how, the Santa Clara Principles should be amended to include broader and more granular requirements for the "Numbers" section. There were plentiful suggestions, and they are set forth below. The countervailing concerns for privacy and competition are addressed in the Recommendations section.

The feedback the coalition received fell into eight categories: government influence on content moderation, transparency around individual content moderation decisions, more information about the use of automated tools, information to help identify trends in content moderation abuse and discriminatory practices, information to help identify systemic trends and effects, and improving transparency reporting practices.

a) Government Influence on Content Moderation

The original Santa Clara Principles are minimum standards for transparency and accountability around companies' enforcement of their own Terms of Service. A number of stakeholders expressed a strong need for the revised Principles to directly confront

the troubling role state and state-sponsored actors play in shaping the companies' content moderation policies and practices government, through takedown requests and other actions. Currently, some internet platforms publish separate transparency reports outlining the scope and scale of government requests for content removals they receive. However, the granularity and consistency of this data varies from region to region, and many stakeholders noted that government cooperation is not always clearly disclosed in these reports. Stakeholders stated that companies should be required to explicitly state any form of cooperation they have with governments and indicate when decisions were required by national and local laws.

Some of the metrics and data points suggested under this category include aggregate data on:

- The total number of requests for content removal received from governmental entities, broken down by country.

- The number of posts removed as a result of government requests for content removal, and the legal or Terms of Service basis for the removal.

- The number of accounts removed as a result of government requests for content removal, and the legal or Terms of Service basis for the removal.

- The number of posts and accounts that have been geo-blocked or restricted by geography in a particular region for violating local laws, and the legal or terms of service basis for that action.

This data should be broken down by the legal reasoning/basis used to justify requests and information on which government agency submitted the requests, including the existence of any court orders.

b) Transparency around Individual Content Moderation Decisions

Numerous stakeholders outlined that although several internet platforms currently publish aggregate information on the scope and scale of their content moderation efforts, there is still a fundamental lack of transparency and accountability around individual content moderation decisions. As a result, these stakeholders recommended that the Principles be amended to encourage companies to publish metrics such as:

- The number of times a post was viewed before it was removed. Stakeholders emphasized that the number of views a post received before it was removed, or its virality, is important for understanding the impact a piece of content had before it was removed.

- The timeline for content removal. This includes data on:
    - Time between when a post was published and when it was removed

- Time before a post that was erroneously removed was reinstated (either as a result of an appeal or as a result of proactive identification of an error by a company)
- Time between user flagging and response from platform
- Time for a piece of content to be identified by an automated tool or the company and then removed
- Time for a piece of content to be flagged to a company and removed by an automated tool.

c) Appeals and Content Moderation Errors

Many submissions underscored a desire for platforms to publish more data on appeals and proactive recognition of content moderation errors in order to paint a more granular picture of the integrity and efficacy of a platform's content moderation operations. Some of the data points suggested include:

- The number of appeals the platform received from users
- The number of appeals received by governments
- The number of/percentage of successful appeals that resulted in posts or accounts being reinstated
- The number of/percentage of unsuccessful appeals
- Time the platform took to review appeal requests and make a decision
- The number of posts the company reinstated after proactively recognizing they had been erroneously removed
- The number of accounts the company reinstated after proactively recognizing they had been erroneously removed
- Breaking down the information on appeals received and reinstatement of content and accounts by content policy violated.

d) More Information About the use of Automated Tools

Many submissions touched on the need for internet platforms to provide greater quantitative transparency around the role automated tools play in their content moderation efforts, among other reasons, to raise awareness among users of the use of AI in content moderation.

- "Although automation technologies are crucial for content moderation, one of its greatest issues is that, in many cases, there is no accurate information about the period, amount, and type of content that has been removed through automated decision making." - *Laboratório de Políticas Públicas e Internet* (Brazil)

At the same time, some respondents also noted that while we need more transparency around automated tools, we should not frame their use as inherently bad or problematic, and should recognize that some use of automation for the purposes of reviewing content was inevitable and sometimes necessary.

Numerous stakeholders recommended that companies publish data outlining the scope and scale of automated moderation efforts, and inform users which specific categories of content or rules violations are targeted by automated systems. They recommended that companies publish data on the amount of content that has been filtered pre-publication, and how much content was prevented from being uploaded. They also recommended that companies preserve all filtered content so researchers can access it. In particular, they want to see transparency around when automated processes are used for content moderation purposes and for what types of content; the criteria that are used by the automated process for making decisions; and the confidence/accuracy/success rates of the processes, including changes over time. In practice, much of this transparency would need to be done through qualitative information rather than quantitative data, although some transparency around the latter (e.g. accuracy rates over time) might be possible. It would also be important for any transparency requirements to recognise that humans are often also involved in content moderation which involves automated processes. Here, however, the respondents simply wanted to know whether ranking (whether downranking or promotion) was used at all and, if so, for what purposes. In practice, most if not all of this transparency would need to be provided through qualitative information rather than quantitative data. A distinction would need to be made between downranking which occurs due to the content itself (e.g. for coming close to a breach of the content moderation policy) and downranking/promotion as a result of data related to the user.

Some of the metrics that were recommended for inclusion in the revised Principles include:

- The number of actions prompted by automated tools, broken out by content policy violated

- Number of posts removed by automated tools and subsequently reinstated as a result of an appeal and/or a company's proactive recognition of an error, also broken down by content policy

- Numbers reflecting the confidence / accuracy / success rates of the processes, including changes over time and whether these rates are improving; these rates could be broken down into more granular categories such as region, language, and type of content (video, audio, text, etc)

- Numbers around algorithmic curation efforts: Companies should publish data on the amount of content that has been filtered pre-publication, and how much content was prevented from being uploaded. It was also recommended that companies preserve all filtered content so researchers can access it.

e) Information That Will Identify Trends in Content Moderation Abuse and Discriminatory Practices

As noted above, many stakeholders voiced concerns that marginalized and minority groups are often disproportionately affected by content moderation harms. Abusive flagging and removal of these communities' content was especially a concern. In order to identify and understand these trends, stakeholders recommended that companies publish data that will enable researchers and civil society groups to identify patterns in which users and communities are being targeted including through abusive flagging patterns, and by whom, and to suggest ways platforms could modify their practices to address them. This data includes:

- Total number of flagged posts
- Total number of flagged posts removed
- Total number of posts flagged by users
- The volume of flags received by region
- The number of heavily flagged posts reported to authorities
- The number of flagged posts categorized by the company as low priority
- The number of flagged posts categorized by the company as not violating their rules
- How much time it took for the company to make a decision on a piece of flagged content after it was reported to them
- The demographics of users who flagged content
- A list of users who flag the most content
- The number of flaggers who have been identified as participating in troll armies
- The number of bot accounts the platform has identified. Bot accounts are often responsible for spreading propaganda and disinformation that targets specific vulnerable groups.

- How specific groups are targeted in a discriminatory manner, for example, through hateful content, violent content, or disinformation.
- The language that removed posts are in
- The country that users who have had their content or accounts moderated live in
- Heavily flagged accounts
- The number of removed posts and accounts whose content policy violations were also violations of human rights (e.g. freedom of speech, the right to privacy, non-discrimination)

Further, several consultations and submissions outlined the need for data points to be adapted based on geography in order to elucidate certain patterns in abuse of the content moderation system. For example, commenters suggested that companies publish the content policy that a flag asserts a piece of content violated and adapt these metrics based on the region. In the United States, categories focused on hateful content, discriminatory content, and gender-based violence should be adapted to include a breakdown of data based on race. In other regions, such as India, the categories should be adapted to allow a focus on caste. One stakeholder noted that these kinds of localized data points are important and that the Principles "should encourage disclosures of information that is not just directly relevant for users but also beneficial in [the] larger public interest of general policymaking."

Some of these suggestions, as well as suggestions below, reveal a tension between the benefits of analyzing personal information about the users of these services and the significant privacy concerns related to the collection, retention, and further use of such information. The revised Principles must be sensitive to not encourage the otherwise unwarranted collection of user data, and must steadfastly avoid urging online services to collect user data as a condition to providing services. However, when a company nevertheless persists in the collection of personal user data, aggregate reporting of that data can be helpful to identify discriminatory trends in content moderation.

f) Reporting Data to Assist in Identifying Systemic Trends and Effects

Many stakeholders expressed concerns that flaws and errors in the content moderation system often disproportionately affect already vulnerable and marginalized communities, and that moderators and systems administrators often lack cultural competence. As a result, they suggested that platforms publish contextual data that can promote understanding of who was affected by a platform's content moderation efforts, when they were affected, where these individuals are based, and why and how they were affected. The data points that were suggested by contributors who focused on this theme are broad, but generally advocated for data on:

- The category of user' appeals that were the most successful (e.g., user, government, etc.)

- The geographic origins of appeals the company has received
- Information about the user whose content was removed, such as demographic background, geographic location, and occupation (e.g., journalist, activist) of users who have had their content removed
- The thematic classification of a removed post and its level of viewership
- The number of times inappropriate content was shared before being removed
- The number of repeat content violations of a given user
- The language and location of the moderator compared to the post and its author, as well as the countries in which content moderation decisions are made
- The languages in which content moderation decisions were made compared to the original language of the post being reviewed

Facebook specifically noted information that they considered should not be included through transparency, namely the quantity of content removed by a specific mechanism of the overall content moderation system (i.e., a specific type of technology or means of review) unless such data addresses a question not capable of being answered by any other metric. Facebook also noted that their content review process is conducted by a combination of people and automation which complement each other, so a piece of content could be flagged by automated tools and then removed by a human reviewer etc. In these circumstances, it would be misleading to suggest that this content was solely moderated by humans or by an automated process. GitHub recommended that any principles developed focus on content moderation and not include other decisions about content display and distribution.

g) Improving Transparency Reporting Practices, While Acknowledging Scale

Several stakeholders also put forth recommendations on how company transparency reports can be improved in order to provide greater visibility into and accountability around corporate content moderation processes and in order to generate meaningful transparency. At least one submission suggested that companies should publish transparency reports on a more frequent basis. Currently, many platforms publish reports on a quarterly basis. Others suggested that companies provide more information about AI technologies used in content moderation regarding the type of content (e.g., video, image, text) and further detail on human intervention in interaction with automated decisions. Commenters also mentioned the relevance to provide more information about which content moderation decisions encompass human intervention and the steps, procedures, and companies involved in the review process:

- Facebook/Meta urged that every metric in a company's transparency report should answer or inform on a critical question connected to transparency, as a form of self-evaluation and a basis for discussion. This was noted as important for generating meaningful transparency. Facebook suggested that the Principles "should seek to lay out core *transparency areas and corresponding questions that transparency can help inform or answer*." Based on this, the Principles can "articulate specific transparency requests and best practices."

Many stakeholders underscored the need for the Principles to encourage and reflect the elements above adding that the Principles would be most useful if they focused on end results instead of breaking down data into overly granular pieces. A participant in the Americas consultation opined that the Principles place too much of an emphasis around quantitative transparency and not enough emphasis on qualitative transparency, as companies often use numbers to justify how well they are doing in terms of content moderation, but the qualitative points provide important context.

h) Numbers related to crisis and emergency-related data

Companies should create a separate category when reporting numbers to outline data related to content removals and restrictions made during crisis periods, such as during the COVID-19 pandemic and periods of violent conflict.

2. Reflections and Observations

The Santa Clara Principles were initially drafted to provide a minimum set of standards that companies should meet in order to provide adequate transparency and accountability around their content moderation efforts. The metrics included in the Numbers section were similarly intended to provide a baseline for what a good transparency report on Terms of Service enforcement would look like. These metrics were released in anticipation of the first-ever industry Terms of Service enforcement reports, which were published by YouTube, Facebook, and later Twitter in 2018. One key question we grappled with was whether the new set of Principles should similarly provide a baseline of standards that companies should meet, whether they should outline a more advanced tier of requirements, or something else. The former would likely allow the coalition to conduct more advocacy with smaller platforms and services that do not currently publish Terms of Service enforcement reports. The latter would put pressure on companies to demonstrate more transparency and accountability in response to changes in the content moderation landscape that have occurred over the past three years.

The metrics and data points noted above all reflect a desire for companies to provide more transparency and accountability around how they design and implement their content moderation systems and what impact these policies and practices have.

However, there are at least three countervailing concerns to more detailed transparency reporting.

First, as some stakeholders recommended, the reporting requirements in the Numbers section may need to be proportional to a platform's level of content moderation, its risk profile, and its size, the platform's age, revenue, and geographic spread. Additional expertise may be needed to fully comprehend the implications of such a decision.

Second, some of the metrics suggested would require the unnecessary and undesirable collection of user's personally identifiable information (PII). For example, many platforms assert that they do not currently collect demographic data such as race or other potentially sensitive information such as location data. Reporting metrics such as these would require platforms to begin collecting this data. In addition, some of the metrics suggested by stakeholders would require the unlawful disclosure of PII. These include metrics that require platforms to identify a user or moderator's demographic information. Alternative methods, such as aggregate or qualitative reporting, for obtaining the information proposed by stakeholders without requiring PII collection or disclosure would have to be identified. Asking internet platforms to provide more information around how the PII they collect factors into content moderation decisions may be useful since many platforms assert that all PII they collect is essential for their operations.

Third, some stakeholders expressed concerns around pushing companies to report on metrics related to timelines for content removal. Over the past several years, several governments have noted that platforms have not been removing content quickly enough. Some of these legislators have subsequently introduced legislation requiring companies to moderate content along a specific timeline, or face penalties. While we recognize the importance of quick action and encourage platforms to moderate content on a timely basis, requiring companies to act along a specific timeline could result in overbroad removals, therefore raising significant free expression concerns. Because of this, the group chose not to include timeline-related metrics in the redrafted Principles.

3. Recommendations

★ As revised, the new Numbers Principle requires that companies report data that reveals how much actioning is the result of a request or demand by a state actor, unless such disclosure is prohibited by law. Companies should also report the basis for such requests or demands—whether a violation of law or terms of service and, if so,which specific provisions, and identify the state actor making the request. Companies should also report the number of requests for actioning made by state actors that the company did not act on.

★ As revised, the new principles also acknowledge the role states play in supporting and promoting transparency in content moderation. States should remove the barriers to transparency that they erect, including refraining from banning companies from disclosing state requests and demands for actioning.

★ As revised, the new principles require that intermediaries report the number of appeals the platform received from users, the number of/percentage of

successful appeals that resulted in posts or accounts being reinstated, and the number of/percentage of unsuccessful appeals.

★ As revised, the new principles require that companies report the number of posts the company reinstated after proactively recognizing they had been erroneously removed, and the number of accounts the company reinstated after proactively recognizing they had been erroneously removed.

★ As revised, the new principles require greater transparency around the use of AI and automated processes in content moderation, including information both about a company's confidence in its automated tools generally and disclosure of all tools used in any specific actioning.

★ As revised, the new principles include a special emphasis on transparency around flagging. The revised Numbers Principle encourages reporting data that will allow users and researchers to assess the frequency of flagging abuse and the measures a company takes to prevent it. Specific metrics and/or qualitative reporting could be devised to help identify abuse-related trends in particular regional contexts, including total number of flagged posts; total number of flagged posts removed; total number of posts flagged by users; the volume of flags received by region; and the number of heavily flagged posts reported to authorities. Companies are encouraged to assess and adapt metrics to regional/local contexts in terms of providing useful information to identify trends in content moderation abuse and discriminatory practices.

★ As revised, the new principles require companies to break down the disclosed data by country and by language, as well as by content policy violated. Companies should provide similar granularity of data across different countries, so as to achieve a greater global consistency in transparency reporting, but also be mindful of specificities of the local context.

★ As revised, the new principles require companies to provide information about the geographical/country and language distribution of content moderators.

C. Notice Principle

1. Summary of Comments

The comments received focused mainly on two areas:

1. Expanding what is in the notice provided to users who post actioned content, and
2. Expanding who receives notice

a) Improving Notice Given to Actioned Users

1. Notice about the Origin of the Flag

The original version of the Santa Clara Principles already recommends that notices indicate whether action against the user was taken as a result of a flag and whether the flag originated with a government. However, recognizing competing data privacy concerns, it does not require platforms to identify flaggers by name if they were non-governmental individuals. Participants emphasized the importance of receiving more information about the source of all flags multiple times throughout the consultation responses. It was particularly noted by respondents outside the US and EU. But generally commenters agreed that the Notice principle should not require a company to reveal private information regarding other users.

Regarding notices about law-enforcement or government flags, commenters recommended including specific information about court orders and other legal bases and authority provided to the company by the officials, even if the purported authority for the removal was the company's own Terms of Service.

However, some respondents raised concerns about overly complex notices providing less useful or overwhelming information to users.

2. Provide users adequate information to support appeal

Several commenters suggested that the principle specify the minimum amount of information the companies should provide to users about the takedown so as to enable a meaningful appeal. This includes screenshots and URLs related to the content violation.

- "When reporting about a content violation, a new mechanism must be attained that deals with sending out all the screenshots and urls related to the content violations. It is not feasible to search for tens and hundreds of urls belonging to the persons profiles involved." - *7amleh* (Palestine)

3. Increased Emphasis on Timeliness

Commenters also suggested that the Notice principle emphasize the importance of prompt and timely notices to enable users to seek meaningful redress. Commenters explained that providers should publicly articulate timelines they intend to meet for providing notice and consider giving pre-action notice in some cases. And users should receive more urgent notifications for more significant post/account restrictions.

- "Information could include the maximum amount of time afforded to lodge an appeal." - *Lawyers' Hub* (Kenya)

GitHub suggested that the principle also acknowledge that there are times when it is not appropriate to provide notifications and describe a set of guidelines for when a provider

can or should decline to provide more detail, such as spam, botnets or legal prohibition on notice itself. As GitHub explained, providing a detailed response in these situations can be counterproductive, particularly because these users may not otherwise use their accounts again.

Also, the Notice principle could specify how to provide notifications to users, with specific recommendations that notifications be sent by at least two different methods (e.g. in-app + by email).

b) Expanding Those who Should Receive Notice

Several commenters suggested that the principle require that users other than the actioned user receive notice or at least some other type of public-facing communications, beyond the notifications provided to users when their content is restricted.

- The public should be notified when formerly publicly-available content was removed and apprised of the reasons for the removal by placing a notice where the content formerly resided.

- Administrators and all members of a group should be informed about any action to suspend or downrank a group.

- Flaggers should also receive notifications about the results of their reports.

2. Reflections and Observations

As with other principles, the recommended additions to the Notice Principle raised concerns around the asymmetric application of the new principles and to avoid generating standards that are impossible for small and medium-sized enterprises (SMEs) to meet compared to large companies.As noted above, the new principles are thus no longer framed as minimum standards but rather mean standards against which any company's practices may be compared.

3. Recommendations

★ As revised, the new principles require that the Notice be in the language of the original post, or at a minimum in the user interface language selected by the user.

★ As revised, the new principles include further recommendations about the information that should be provided to users when a state actor is involved with the flagging and removal of the user's post or account.

★ As revised, the new principles specify the full range of actioning for which notice is required.

★ As revised, the new principles specify that companies should provide information about the removal of a post or account at the content's original URL (tombstoning).

★ As revised, the new principles require that notice be timely and that users be notified of any time limitations on appeals procedures.

★ As revised, the new principles specify that any exceptions to the Notice Principle, for example when the content amounts to spam, phishing or malware, be clearly set out in the company's rules and policies.

★ As revised, the new principles require notice to users other than the author of the post, including group administrators, flaggers, and in some circumstances, the public.

D. Appeals Principle

1. Summary of Comments

The comments we received generally sought to advance one of two goals: (1) to ensure due process in the manner in which appeals are considered, and (2) to yield reportable data for research and systemic accountability. And as with the comments on the other principles, many concerns were raised regarding cultural competency.

a) Procedures to Ensure Due Process in Appeals

1. Clarify the definition of "appeal"

We received several comments about the need to further define what is meant by an "appeal." Does the principle require a formal process? Should it require a panel rather than a single decision-maker? Should the appeal be heard by a certain authority?

Several comments sought clarification of the phrase "not involved in the initial decision," with one commenter suggesting that the principle specify that the appeal be heard by people "not in the same chain of command" as the original decision-makers.

- "'Person or panel of persons not involved in the decision' is a good minimum baseline, but ideally the appeal process would involve people not in the same 'chain of command' so to speak." - *Public Knowledge* (US)

2. Clarify the scope of appeals

Several comments pertained to expanding or narrowing the scope of appeals.

With respect to broadening the scope of appeals, one commenter suggested that appeals be available to those users who request that abusive content be removed, but whose request is denied, and to users who flag content for other reasons but where no action is taken. Others suggested that decisions other than takedowns and account suspensions be appealable. These included downranking and shadowbanning.

- "We recommend that this principle be expanded to provide further consideration for the targets of online abuse. Based on our extensive work with artists, writers, and journalists globally, a significant pattern of abuse involves the weaponization of takedown mechanisms against these groups, the expediency of appeals is often critical in these cases because the timeliness of the messages that are targeted is often critical. For example, takedowns are frequently wielded against activists during protests and during elections. The eventual restoration of these kinds of content does not ameliorate the efficacy of these tactics in disrupting communication and expression in real time.

  "We additionally recommend that the principle be expanded to include the need for an appeals process for those who report abusive content but are informed that it does not violate terms of service - or receive no response at all. In these cases, we recommend that a second review can be requested, that the reporting party be provided with additional information regarding how the case was evaluated, and that the reporting party be given the opportunity to provide context on the content they have reported." - *PEN America* (US)

With respect to narrowing appeals, GitHub, a platform that has to make such decisions, suggested that where a user is contesting the basis of a decision, a platform not be required to consider appeals unless the user provides additional information regarding the facts of their dispute.

- "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 new information for them to consider." - *GitHub* (US)

GitHub also suggested that the principle allow for exceptions for removals or suspensions of spam, phishing, or malware in certain circumstances, such as where there is large-scale, apparent malicious activity or intent.

- "We recommend defining "appeal" and allowing for exceptions in certain cases, for example, for spam, phishing, and malware." - *GitHub* (US)

Others agreed that some takedown decisions deserved greater appellate consideration than others. For example, a more robust appeal mechanism may be required for removals based on editorial and other subjective judgments, in contrast to those based on the application of reasonably objective rules.

- One participant in the European virtual consultation said that it would be wise to have more guarantees in terms of appeal for a person whose post was removed, allegedly from being fake news, and then fewer appeal opportunities and guarantees for a person whose post was removed for allegedly constituting child pornography. "I wonder if it could be possible to distinguish between removals that are applications of reasonably objective rules (e.g. certain images not allowed no matter what) vs. those that involve editorial discretion." - *Public Knowledge* (US)

3. Allow users to submit additional information

Several commenters requested that the principle specify information that users should be able to submit to support their appeal. This should include information about takedowns and suspensions from other websites, similar patterns of content posting, and any evidence the user has regarding discriminatory targeting.

- "Further specificity should be given with regards to the additional information that may be provided during appeal. Such information should be specified to include, at a minimum, content takedowns and/or account suspensions from other websites, similar patterns of content posting and where possible, political affiliations to groups and/or organisations deemed to be discriminatory." - *Lawyers' Hub* (Kenya)

4. Appeal procedures must be clear

Several commenters suggested that the principle require that the appeal procedures be clearly set out and easy to understand, with one commenter suggesting that the process be illustrated with visuals.

- "Platforms should provide clear guidelines on the appeal process, as well as data on prior appeals. . . . The process should also aim to be something that a lay individual can complete without having to seek external legal counsel as that itself might become a barrier for people to appeal decisions that they deem to be unfair if they don't have the means to access services that can help them appeal those decisions. Following a GDPR-like requirement for making the legalese accessible and understandable to the users, we propose that platforms also invest in handy tutorials, both text and video, that make it easier for people to complete the process on their own." - *Montreal AI Ethics Institute* (Canada)

Other commenters focused on the need for specific information regarding appeal timing and deadlines—the time available to file an appeal and the time a user should expect for the appeal to be resolved—and mechanisms to allow users to track the progress of appeals. And users should be promptly notified of the decision on the appeal and the consequences and finality of that decision.

- "Users should be afforded access to an application that allows them to track the status of their appeal throughout the process and offers benchmarks for timeliness of the stages of review, as well as recourse in the event that the process drags out." - *PEN America* (US)

5. Procedures for expedited appeals

Some commenters noted that in cases of crisis and urgency, an appeal may be truly timely only if it is expedited and considered on an as soon as possible or emergency basis. The principle should thus require that users be able to request such expedited consideration.

6. Cultural competence in appeals

Numerous commenters would require the principle to ensure cultural competence in appeals processes. Many would define "meaningful opportunity for appeal" to include reviewers with language fluency and cultural and regional knowledge of the context of the post, that appeal procedures be available in the language of the users, and that assistance be available during business hours in the user's time zone. Responses to appeals should be issued in the language of the original post.

- "They should also make clear what kind of teams take care of the appeal process and how they deal with the culturally and linguistically specific circumstances." - *Montreal AI Ethics Institute* (Canada)

- A participant in a consultation held in Africa suggested that appeals processes should be easily understandable to the ordinary user, and be available in different languages and time zones. The participant suggested that users be able to respond to notices and lodge appeals in a language they feel comfortable communicating in and that Policies should be available in multiple languages so that users are aware of and can understand community standards of different platforms.

7. Procedures need to show sensitivity and responsiveness to abuse of takedown schemes

Several commenters suggested that the principle requires that appeal procedures show special sensitivity to users who have been subject to abusive targeting, or whose content removals are part of a broader scheme of online abuse.

8. Clearly explain appeal results and consequences

Several commenters also suggested that the principle require the platforms to clearly explain the results of the appeal and any available remedies. One commenter suggested that platforms allow users the opportunity to revise a removed post to address noted community standards or terms of service violations. Another suggested that platforms make remedies available to address damages caused by reversed takedowns or wrongly rejected takedown requests.

- "If there is in fact a ToS violation, users should have the opportunity to revise their post to have it reposted." - *Association for Progressive Communications* (International)

9. Need for external review

Some commenters suggested that external review be required in certain cases, or at least to assess efficacy of the internal review processes. This was a special concern for oversight of automated decision-making. An external reviewer or ombudsperson might be able to detect larger trends in how rules are enforced rather than seeing a moderation decision in a vacuum of the platform's numerous other decisions. The Facebook Oversight Board was given as an example.

- "Currently, the category of 'appeals' in the Santa Clara Principles is focused on having accountability processes in place, and emphasizes the need of having meaningful review. The framework of the Principles also currently envisage only internal review processes carried out by the company. However, in light of Facebook unveiling its plans for an Oversight Board, a structurally independent body, which would arbitrate select appeal cases of content moderation, these pre-existing principles might need revisiting." - *Centre for Internet and Society* (India)

b) Yield Data to Further Research and Systemic Accountability

The second category of comments addressed the need for appeals to yield data to facilitate research and external oversight of the platform and the larger online information ecosystem. This data will allow researchers to assess the accuracy of content moderation decisions and the effectiveness of appeals. This includes data on both successful and unsuccessful appeals

Several commenters suggested that the principle require platforms to report data regarding the results of the appeals; for instance, how many were successful and how many were unsuccessful, as part of larger data collection to assess overall efficacy of content moderation systems.

- "We recommend that the following additional data sets be collected for the purposes of evaluating the effectiveness of content moderation among specific companies:

  "The number of successful appeals. This would serve as an indication of the mechanisms that are currently being employed for flagging content and provide information on the areas upon which they can be improved.

  "The number of unsuccessful appeals. This data would indicate the level of certainty and familiarity that users have with regards to the rules used on online platforms.

  "The average speed of detection of inappropriate content. This would enable various companies to determine whether they are detecting flaggable content in a timely manner and therefore preventing its spread on their platforms.

> "The amount of times inappropriate content was shared and/or viewed before being flagged and taken down; Recording such data would allow companies to establish the means of the spreading of flagged content as well as to determine the damage and harm that is caused by the spreading of said content.
>
> "The number of repeat offences. This would enable various companies to determine the perception of the rules and regulations concerning content moderation among offenders and help them model them in a manner that would deter future violations." - *Lawyers Hub* (Kenya)

Commenters also suggested that platforms provide information that will assist future users in developing their appeals. This would include information about what a successful appeal was dependent upon—the quality of the submission, executive intervention, etc.

2. Reflections and Observations

Many commenters found Appeals to be the most complete component of the existing Santa Clara Principles and there were far fewer comments than for Numbers and Notice.

Over time, platforms have expanded the ways in which they moderate content beyond removals and account suspension. Now, enforcement actions range from the application of labels to downranking to demonetization. The ability to appeal these new types of moderation has failed to keep up. In some situations, such as demonetization or the application of warning labels, expansion of appeals may be relatively straightforward. In others, such as downranking or shadowbanning, users may not even know that such measures were applied. Platform ranking algorithms make a number of decisions that impact the placement of particular pieces of content; these decisions may be fundamentally different from decisions made through a traditional flagging process.

3. Recommendations

- ★ As revised, the new principles include a specific reference to the importance of cultural competence among staff conducting the review, and require that appeals be considered in the user's language by those with cultural understanding of the post's context.
- ★ As revised, the new principles specify that users be able to submit additional information in support of their appeal.
- ★ As revised, the new principles specify the full range of actioning that should be appealable, and consideration should be given as to whether certain types of actions require more or less rigorous appeals processes.

★ As revised, the new principles require companies to provide a meaningful opportunity for timely appeal of decisions to remove content, keep content up which had been flagged, suspend an account, or take any other type of action affecting users' human rights, including the right to freedom of expression.

★ As revised, the new principles reflect the importance of providing users with information about their access to any independent review processes that might, and confirm that any such independent review process should also embrace the Santa Clara Principles and provide regular transparency reporting and clear information to users about the status of their appeal and the rationale for any decision.

★ As revised, the new principles direct companies to consider whether certain targets of abusive takedown schemes should by default be entitled to expedited appeals in certain situations.

★ As revised, the new principles require that appeal procedures be clear, complete, and understandable, including specifying all deadlines. Companies should also develop systems for users to track the progress of appeals.

★ As revised, the new principles require the periodic reporting of appeals data, whether to the public or to researchers.

E. Advertising and the Santa Clara Principles

We asked respondents whether the Santa Clara Principles should have special provisions for how advertisements are served to users, and otherwise incorporate advertisements into the Principles.

The overwhelming majority of respondents were supportive of the proposal to extend the principles to include recommendations around AI-based advertisement targeting and delivery. On the question of whether they should apply to the general moderation of advertisements, around half of respondents simply answered "yes" to this question, suggesting that the Santa Clara Principles should apply in full to the moderation of advertisements. The other half, however, suggested that while there should be greater transparency over the moderation of advertising, the existing Santa Clara Principles would need adaptation. This was justified on the basis that different principles and considerations applied when it came to how platforms develop policies and moderate their commercial and paid-for advertised content as opposed to user-generated content.

A number of suggestions as to what types of information should be required when it came to the moderation of advertising generally were put forward, including:

- Information on how advertisements are clearly labelled and identified as opposed to other pieces of content;

- Information on the criteria by which advertisements are targeted;
- Information on how much money was paid for each advertisement; and
- Information on the source of payment for advertisements, especially political advertisements.

Commenters did not have many specific suggestions about what transparency should look like when it came to AI-based targeting and delivery of advertisements; the two most common themes were the need for information about the types of data collected from users for advertising purposes; and information on how users are categorised or segmented for advertising purposes.

Further suggestions for information made by a smaller number of respondents were:

- How users are categorised or segmented for advertising purposes;
- Whether data collected for advertising purposes is used for other purposes and, if so, what those other purposes are;
- What notice, if any, users are provided with when their data is collected for advertising purposes; and
- The broader policies of the platform as to what types of advertisements are allowed and not allowed.

Facebook specifically suggested that any transparency relating to advertising (or "paid content") through the Santa Clara Principles should be framed to take account of the special characteristics of the advertising ecosystem. Participants in the Latin America virtual consultation highlighted that companies should aim for greater global consistency in their transparency initiatives, although platforms' advertising policies in each country may be affected by local law.

Reflections and Observations

★ The overwhelming majority of respondents also want to see more **transparency around the use of AI and automated processes in advertising targeting**. Here, the most common requests were for the Santa Clara Principles to include requirements for information on the types of data that are collected from users for advertising purposes, and how users are categorised or segmented for advertising purposes.

★ The overwhelming majority of participants also wanted to see more **transparency around the moderation of advertisements**. Many respondents noted that there were differences between paid content and unpaid user-generated content which required different considerations. Despite this,

around half of respondents in support of greater transparency felt that the existing Santa Clara Principles were sufficient, while the other half felt that they needed adaptation. While political advertising was seen by many as a particularly important focus, recommendations on transparency relating to advertising generally included requiring more information on how advertisements are clearly labelled and identified as opposed to other pieces of content; the criteria by which advertisements are targeted; how much money was paid for each advertisement; and the source of payment for advertisements, especially political advertisements.

Recommendations

- ★ As revised, the new principles require that its mandates apply equally to paid content as to unpaid user content. As stated in the new principles: **"The term "content" refers to all user-generated content, paid or unpaid, on a service, including advertising."** The following were considered for inclusion in the revised Santa Clara Principles with respect to the moderation of advertisements: the source of advertising, the amount paid for advertising, required labeling of advertising, targeting criteria selected by the advertiser, targeting criteria selected by the service, differential rules for different types of advertising, such as electioneering, whether to require equivalent transparency in relation to advertising content as compare to unpaid user-generated content. Respecting differences in legislation and/or other local particularities, companies should provide similar information and data points across the countries they serve ads.

- ★ As revised, the new principles include greater transparency around the use of AI and automated processes in advertising targeting.

- ★ The revised Santa Clara Principles emphasize that cultural competency is required for the moderation of advertisements to the same extent it is for unpaid user-generated content.

Acknowledgements

Thank you to all of the organizations and individuals who submitted comments, participated in the group consultations, and reviewed and commented on preliminary work. Organizations submitting comments include the following: 7amleh, Association for Progressive Communications, Centre for Internet & Society, Facebook/Meta, Fundación Acceso, GitHub, Institute for Research on Internet and Society, InternetLab, Laboratório de Políticas Públicas e Internet (LAPIN), Lawyers Hub, Montreal AI Ethics Institute, PEN America, Point of View, Public Knowledge, Taiwan Association for Human Rights, The Dialogue, Usuarios Digitales. The list of individuals and groups who coordinated and hosted consultations, and otherwise contributed to the process, includes, but is not limited to: ALT Advisory, Centro de Estudios en Libertad de Expresión y Acceso a la Información (CELE), UNESCO, Irina Raicu, Eduardo Celeste,

Derechos Digitales, Robert Gorwa, Ivar A.M. Hartmann, Amélie Heldt, Tomiwa Ilori, Julian Jaursch, Clara Iglesias Keller, Paddy Leerssen, Martin J. Riedl, Christian Strippel, and Daphne Keller.