The Santa Clara Principles on Transparency and Accountability in Content Moderation: Public Consultation

Submission by the Association for Progressive Communications (APC)[1]
June 2020

Introduction

APC is an international network of civil society organisations founded in 1990 dedicated to empowering and supporting people working for peace, human rights, development and protection of the environment, through the strategic use of information and communication technologies (ICTs). We work to build a world in which all people have easy, equal and affordable access to the creative potential of ICTs to improve their lives and create more democratic and egalitarian societies.

As an organisation that has worked at the intersections of human rights and technology for nearly three decades, we welcome this public consultation on the Santa Clara Principles on Transparency and Accountability in Content Moderation, as it is timely and integral to our work. Censorship is increasingly being implemented by private actors, with little transparency or accountability, and disproportionately impacts groups and individuals who face discrimination in society – groups and individuals who look to social media platforms to amplify their voices, form associations, and organise for change. The current pandemic has raised challenges for content moderation. While we recognise that these are extraordinary times, human rights laws and principles should be the default standards guiding companies' content moderation policies and procedures.

[1]We thank APC individual member Damián Loreti for providing comments during the development of this submission.

In general, the number of platforms releasing transparency reports and the details contained in those reports has increased in recent years. Companies, however, still need to do more to comply with their responsibility to ensure that their policies for restricting content are being applied in a non-discriminatory and equitable manner. Platforms report very little about when they remove content or restrict users' accounts for violating their terms of service (ToS), and often fail to provide adequate notice: users are not adequately informed about the rules they have violated. Moreover, companies enter into agreements with states to operate locally, and the terms of these agreements, which have implications for content regulation, are often completely unknown.

We encourage the Principles to urge companies to improve reporting mechanisms following these criteria:

- **Legitimacy:** The mechanism is viewed as trustworthy and is accountable to those who use it.
- **Accessibility:** The mechanism is easily located, used and understood.
- **Predictability:** There is a clear and open procedure with indicative time frames, clarity of process and means of monitoring implementation.
- **Equitable:** It provides sufficient information and advice to enable individuals to engage with the mechanism on a fair and informed basis.
- **Transparency:** Individuals are kept informed about the progress of their matter.
- **Human rights-respecting:** The outcomes and remedies accord with internationally recognised human rights standards.
- **Source of continuous learning:** It enables the platform to draw on experiences to identify improvements for the mechanism and to prevent future grievances.[2]

1. Currently, the Santa Clara Principles focus on the need for numbers, notice, and appeals around content moderation. This set of questions will address whether these categories should be expanded, fleshed out further, or revisited.

a. The first category sets the standard that companies should **publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines. Please indicate any specific recommendations or components of this category that should be revisited or expanded.**

Transparency reports refer mainly to numbers: those of pieces of content taken down, how many appeals, how many items restored after appeals. The Santa Clara Principles contribute to this understanding by focusing the first set of recommendations on the broad category of "numbers". It would be desirable for the Principles to also include a qualitative approach to this section. This could influence companies to, for example, provide more analytical data on the processes and motivations leading to take-downs, as well as decision-making parameters.

We encourage the Principles to request more transparency around content moderation related to gender-based violence (GBV). GBV is rampant online and has become even more complex in recent years, ranging from hate speech to the non-consensual distribution of intimate images.[3] Within the broad

[2]Various. (2017). *Statement on Facebook's internal guidelines for content moderation*. https://www.apc.org/en/pubs/statement-facebooks-internal-guidelines-content-moderation

[3]Take Back the Tech!, Luchadoras, & SocialTIC. (2018). *13 manifestations of gender-based violence using technology*. https://www.genderit.org/resources/13-manifestations-gender-based-violence-using-technology

category of online GBV, detailed and disaggregated data on the different types of violence would be relevant.

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.

Companies should also disclose information on enforcement of their ToS. They should make public their guidelines on how ToS are implemented. There should be an opportunity for feedback from the community of users and human rights experts to ensure that implementation guidelines comply with principles of equality and non-discrimination.

In general, platforms should provide greater transparency and accountability regarding the following:

- The implementation of content moderation guidelines.
- The rejection of reports of online abuse and disaggregated data on reports received.
- Data on accuracy of human and automated detection.
- The departments and staff responsible for responding to content and privacy complaints.

b. The second category sets the standard that companies should **provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension. Please indicate any specific recommendations or components of this category that should be revisited or expanded.**

Users are not consistently being notified about content restrictions, take-downs and account suspensions, the reasons for such action, and the procedure they must follow to seek reversal of such action.[4]

Companies should give notice to each user whose content is taken down or account is suspended about the reasons for the removal or suspension. The Santa Clara Principles should reinforce that notification should be expedited and users should be notified that their content has been restricted, and informed of the basis on which it was restricted. This should be done using accessible language and referencing which human rights laws and principles were violated. When it is found that reported content is problematic but does not qualify for removal, platforms should find a way to communicate this with language that explains possible rights violations and implications.

c. The third category sets the standard that companies should **provide a meaningful opportunity for timely appeal of any content removal or account suspension. Please indicate any specific recommendations or components of this category that should be revisited or expanded.**

As part of their responsibilities under the UN Guiding Principles on Business and Human Rights, companies are required to provide access to remedy in order to mitigate harm, and grievance mechanisms are critical in this regard. Companies should notify users why content was restricted, taken down, or accounts suspended, and allow them to appeal immediately.

[4]Barbour, M. (2020, 25 May). Cummunity standards. *GenderIT.org*. https://www.genderit.org/feminist-talk/cummunity-standards

If there is in fact a ToS violation, users should have the opportunity to revise their post to have it reposted. The avenues for appealing the decision should be presented in understandable and accessible language. Users should be able to appeal in their own language and reach someone in their own time zone. Without such measures, users may continually and unintentionally repeat the violation and have their account disabled, exacerbating the violation of their freedom of expression.

Platforms need to take action and also look at structural and systemic problems. Companies need a better reporting system that allows for context and stepping back from a post to be able to make informed decisions about malicious use that goes beyond a specific individualised post.

These series of consultations are an opportunity for the Santa Clara Principles to expand so that they encourage companies to educate users, rather than simply taking down content. For instance, users should be informed of repeated posts that carry disinformation that they share. And if there is a systemic pattern, companies should take action. Or, if users are violating ToS through non-consensual sharing of intimate images, for example, notifying them of the reason for the take-down is an opportunity for the platform to educate the user on its anti-violence standards and definitions.

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

Content moderation (the use of artificial intelligence to review content and accounts and determine whether to remove the content or accounts; processes used to conduct reviews when content is flagged by users or others)

Content ranking and downranking (the use of artificial intelligence to promote certain content over others such as in search result rankings, and to downrank certain content such as misinformation or clickbait)

Ad targeting and delivery (the use of artificial intelligence to segment and target specific groups of users and deliver ads to them)

Content recommendations and auto-complete (the use of artificial intelligence to recommend content such as videos, posts, and keywords to users based on their user profiles and past behavior).

These are all emerging issues and the Santa Clara Principles should be expanded on all of them. Regarding the use of artificial intelligence (AI) for content moderation, at APC we believe AI systems could help to identify problematic content for human review, as we are aware of how stressful the work of content moderators who have to spend hours looking at often very disturbing content can be. Automation could be employed to facilitate human moderation at scale. AI, however, should not fully replace human moderation.

When such automated processes are used, their use should be made transparent, all content removal should be subject to human review, and users should have easy recourse to challenging removals which

they believe to be arbitrary or unfair. Information should be disclosed on how AI tools are designed, trained, applied and refined.[5]

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

APC has used the Santa Clara Principles as a reference and resource for several inputs and submissions for United Nations Special Procedures and consultations on content moderation. However, it would be desirable to strengthen the effectiveness and implementation of the Principles. Besides being useful standards, there is a need for more efforts towards greater enforcement and effective use by companies.

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

Building on our previous response, it would be interesting to explore how these principles could be translated into practical implementation. For instance, the Principles could be complemented by a toolkit or guidelines for implementation. Annual reports and discussion meetings could be organized with adhering platforms to discuss compliance challenges and successes.

Another issue that could be raised by the Principles is the periodicity of companies' transparency reports. They are released every five or six months. The Principles could refer to a minimum of when these reports should happen, since they are our only source of information on content takedowns at scale by companies.

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

It is still unclear to us how the Principles are being implemented by companies. It would be great to expand on how the principles work in practice and on accountability around their implementation.

It would be interesting to acknowledge in the text the challenge of fast-changing contexts and indicate how the Principles are planned to be amended and adapted as circumstances and technology evolve, for instance, in terms of more and more reliance on AI moderation systems and less review by humans.

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

It is only through civil society and academic research that it becomes clear who is being silenced or faces more censorship or aggravated online violence.

Women facing violence online, black journalists facing discriminatory comments, women in public positions[6] – these are all examples of those disproportionately experiencing online abuses; these

[5]Singh, S. (2019). *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*. New America. https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/promoting-fairness-accountability-and-transparency-around-automated-content-moderation-practices

[6]Gardiner, B., et al. (2016, 12 April). The dark side of Guardian comments. *The Guardian*. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

categories would not in any way be apparent from the transparency reports currently released by the companies. Of course, concerns with privacy, bias and profiling should be seriously considered, so a delicate balance should be struck here.

It would be important for the Principles to encourage not only transparency that would be directly relevant to users, but also the disclosure of information that would be useful from a broader public interest/policy point of view, including for the protection of human rights online and to provide clearer parameters for the work of legislators and policy makers.

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

Standards could vary, but international human rights standards should be the minimum and should be applied equally, meaning that all users should enjoy the same procedural safeguards and due process irrespective of what jurisdiction they live in or language they speak.

It is worth mentioning that the local context or jurisdiction should influence the understanding on who vulnerable groups are. Contextual analysis of the power dynamics in different jurisdictions is critical to understanding whether certain content constitutes incitement to discrimination, hostility or violence. Arguing for differential standards, however, especially for online spaces where there is already ambiguity around what laws and standards apply, could be discriminatory.

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

A distinction could be made between the global platforms that effectively serve as a gateway for information and an intermediary for expression among populations, and smaller platforms that serve more selective audiences, which might want to tailor their terms of service/community guidelines to their respective communities.

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

Online speech is dynamic, so content moderation rules and policies are inserted into an ecosystem that is constantly evolving. As mentioned before, it would be interesting to insert into the Principles tools aimed at ensuring they remain updated and flexible. Initiatives like this one – to review, adapt and expand the Principles – should be carried out periodically.

This review should include groups most affected by content moderation, civil society and human rights defenders, especially from the global South.

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

It would be desirable to ask for more transparency, overall, around the human moderation of content. Companies should employ moderators from diverse backgrounds in terms of characteristics such as

economics, race/ethnicity, caste, religion, language and region; and should provide adequate training and management.

But we do not believe the Principles should 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. However, it would be interesting to think of the Principles as a tool to also ensure moderators' labour rights.

In addition, more than focusing on moderators, it would be interesting to see how the Santa Clara Principles will dialogue with the Facebook Oversight Board.

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

The COVID-19 crisis has raised challenges for content moderation. Companies have been increasingly relying on automation.[7] Users, in this complex context, have received notices saying that there are no human resources to deal with their complaints.

While we recognise that these are extraordinary times, human rights laws should be the default standards guiding companies' content moderation policies and procedures, and principles of accountability, transparency and meaningful appeal processes, among others, should guide content moderation responses to the pandemic.

Another concern that has become even more explicit during the COVID-19 pandemic is that of the working conditions of human moderators.[8] Relevant information could be made available also in this regard.

It would be interesting to discuss the publication of transparency reports specifically related to the particular measures taken by platforms to address the health crisis, possibly with a focus on actions to tackle the spread of COVID-19-related misinformation.

With less human intervention and more automation, what happened to appeals? As civil society groups have stated, companies should preserve all data on content removal during the pandemic, including but not limited to information about which take-downs did not receive human review, whether users tried to appeal the take-down, and reports that were not acted upon. All content that the platform is automatically blocking or removing, including individual posts, videos, images, and entire accounts, should be preserved.[9] In addition, more transparency around the use of automation is needed, since it has been reported that there has been greater reliance on AI than companies have recognised.[10]

[7]Rosen, G. (2020, 12 May). Community Standards Enforcement Report, May 2020 Edition. *Facebook*. https://about.fb.com/news/2020/05/community-standards-enforcement-report-may-2020

[8]Biddle, S. (2020, 12 March). Facebook Contractors Must Work in Offices During Coronavirus Pandemic – While Staff Stay Home. *The Intercept*. https://theintercept.com/2020/03/12/coronavirus-facebook-contractors

[9]Various. (2020). *Open letter on COVID-19 content moderation research*. https://www.apc.org/en/pubs/open-letter-covid-19-content-moderation-research

[10]Chaturvedi, A. (2020, 12 May). Facebook unveils fifth Community Standards Enforcement Report. *The Economic Times*. https://economictimes.indiatimes.com/tech/internet/facebook-unveils-fifth-community-standards-enforcement-report/articleshow/75702888.cms