The Santa Clara Principles

On Transparency and Accountability in Content Moderation

Toolkit for Advocates

I. History of the Santa Clara Principles

In May 2018, a coalition of organizations, advocates, and academics came together to create the Santa Clara Principles on Transparency and Accountability Around Content Moderation in response to growing concerns about the lack of transparency and accountability from internet platforms around how they create and enforce their content moderation policies. The Principles outline minimum standards that tech platforms must meet in order to provide adequate transparency and accountability around their efforts to take down user-generated content or suspend accounts that violate their rules.

The original set of Principles focuses on three key demands—comprehensive numbers detailing a platform's content moderation efforts, clear notice to impacted users, and a robust appeals process. They are consistent with the work of David Kaye, former UN Special Rapporteur on the promotion of the right to freedom of expression and opinion, who called for a "framework for the moderation of user-generated online content that puts human rights at the very center." The principles also reflect the recommendations of the UN Guiding Principles on Business and Human Rights, which articulate the human rights responsibilities of companies.

When the Principles were first released, there was very little transparency around the scope, scale, and impact of internet platform's content moderation efforts. As a result, the authors of the Principles called on companies to disclose more data around these moderation efforts via transparency reports. These transparency reports have helped highlight government censorship on platforms, enabled users to make more informed decisions about which products to use and avoid, and empowered advocacy groups to push companies to follow established legal processes when responding to and complying with government demands. Additionally, the authors of the Principles noted that content moderation often occurs in a top-down manner, leaving users with few options for remedy and redress. The "notice" and "appeals" Principles sought to establish robust, transparent, and reliable mechanisms for due process for users.

Since their release, many internet platforms have endorsed and committed to adhering to the Principles. These platforms include Apple, Facebook, GitHub, Google, Instagram, LinkedIn, Medium, Reddit, Snap, Tumblr, Twitter, and YouTube. While some of these platforms have made notable strides in providing more transparency around their content moderation efforts, very few companies have fully met the demands outlined in the Principles. Platforms must do more to meet these baseline expectations of transparency and accountability.

The Santa Clara Principles coalition has launched an updated set of principles in order to further platform transparency and accountability.

While the original 2018 Principles set forth very strong baseline standards with which companies should comply, participation in their creation was limited to just a few groups and individuals, and allies—particularly from countries outside the United States and Western Europe—raised legitimate concerns and suggestions for their revision. In particular, stakeholders from around the world have emphasized that platforms are investing more resources in providing transparency and due process to users in certain communities and markets. Companies must address this inequity and ensure that all of their users—regardless of where they live—can obtain transparency and accountability from these companies. This is particularly important given that many of the harms that occur as a result of platform content moderation practices occur in communities that platforms have been neglecting.

The content moderation landscape has radically changed over the past few years. Platforms are no longer tackling harmful content and accounts by simply removing them. Today, many services also rely on algorithmic tools to curate content through interventions such as downranking. There is a serious lack of transparency and accountability around how platforms are deploying these interventions and what the resulting impacts on freedom of expression are. Additionally, researchers and advocates have underscored the discriminatory and harmful outcome that can result from paid content online. There is currently also a major lack of transparency around how such content is moderated, and with what impacts. These are additional areas that platforms must commit to shedding light on.

Lastly, during the COVID-19 pandemic, many platforms shared that they would increase their reliance on automated tools for content moderation purposes. Some services also announced that they would be suspending their appeals processes, therefore impeding users' access to due process. Numerous civil society organizations expressed concerns around how these decisions would impact freedom of expression online, underscoring that platforms must be able to maintain a baseline level of transparency and accountability at all times.

Because of these three concerns, the Santa Clara Principles coalition initiated an open call for comments from a broad range of global stakeholders, with the goal of eventually expanding the principles. The coalition engaged in significant public and community outreach via an open comment period and complementary targeted outreach strategy, then reviewed the inputs during a designated period, and finally, drafted a new set of Principles. A series of open consultations and workshops were held to add more details to the original set of principles.

II. A Toolkit for Advocates

This campaign toolkit seeks to explain the importance of the Principles, key messages, and provide insight into how advocates can campaign independently and as a coalition to hold companies to account—not only to endorse the new Principles but to implement them in their policies and practices.

Tech companies control online information flows on their platforms through proprietary rules and Terms of Service, giving them significant power with little accountability. Communities already facing discrimination are also at risk of having their content removed online through discriminatory flagging campaigns or biased moderation processes, and thus face being doubly silenced.

Wrongful action taken on content can have a disproportionate impact on already-vulnerable populations, such as members of ethnic or religious minorities, LGBTQ+ people, and women. It also routinely affects journalists, political activists, and human rights defenders operating in repressive environments.

Governments are currently looking to regulate internet platforms to ensure that harmful content is removed quickly and steps are taken to prevent such content from appearing in the first place. However, there have been a number of regulatory proposals that put extralegal pressure on social media companies to remove content at a rapid pace or seek to hold platforms liable for third-party speech—effectively ensuring that large platforms will retain their near-monopolies. As companies face regulatory pressure, they are likely to increase the speed of content moderation and their use of automated technologies in order to avoid facing hefty fines. Companies are likely to make more content moderation errors when operating under time-bound pressure. Additionally, automated tools used for content moderation are limited in a number of ways, often resulting in the removal of too little or too much speech. This raises significant freedom of expression concerns.

We need an urgent solution that ensures the Internet is a space for all people to access information and take part in debate. Putting in place clear notice and appeals processes is a basic first step that social media platforms can take to make sure that all users can be heard and to protect online communities.

III. Key Targets for the Principles

A. State actors

State actors must abstain from passing legislation that hinders freedom of speech on the internet and ensure that human rights are being protected on platforms. The Principles are intended for the governments to have some context of the average standards and good practices regarding content moderation online.

The Principles are not intended to be a template for regulation, but a guide so governments know what kind of standards they should take into consideration when discussing regulation or policy. Moreover, state actors must remove obstacles to obtain transparency from companies and also report their involvement in content moderation decisions. Finally, state actors must acknowledge civil society’s important role in promoting freedom of speech online and foster a multi-stakeholder approach to content moderation discussions.

B. Social Media Platforms

Social media platforms have become fundamental to how we communicate. When users join a platform, they agree to that platform's Terms of Service, which typically oblige the user to abide by a set of rules about acceptable behavior and speech, often contained within a separate "community standards" document. Users who run afoul of these standards may find their content actioned*.

Companies are increasingly subject to demands from governments or legislation that holds them liable for user expression, raising serious questions about the future of free expression online. Concerns from global civil society about the lack of transparency and accountability from platforms prompted the initial creation of the Principles, while the changing nature of content moderation and its uneven application globally brought forth new demands and created the impetus for an evaluation and revision of the Principles. The Santa Clara Principles 2.0 provide a new set of standards for transparency and due process that serve as a guide for companies to enact human rights-preserving measures into their policies and practices.

We encourage members of civil society, companies, and other stakeholders to work together to develop implementation plans in consultation with various stakeholders in order to develop a roadmap toward adherence to the revised Principles. We also encourage civil society to hold companies accountable in executing their roadmaps.

III. Using the SCP 2.0 in Your Advocacy

A. Build visibility of the Principles

by holding a webinar or press conference inviting keynote speakers, companies, government and activists to speak about freedom of expression online and key recommendations from the Principles.

B. Organise face-to-face meetings with the companies in your country.

Make sure to bring in activists who have had negative experiences with content takedowns. Putting pressure on companies through evidence-based advocacy is the key! It is important that when meeting the companies, we hold them accountable to what was agreed. Take notes of the discussion and action points during the meeting. End the meeting by requesting a follow up discussion with the companies. One meeting is not enough! Hold companies accountable in the follow up meeting to respond to the action points that were previously discussed.

C. Organise face-to-face discussions with relevant state actors in your country.

Make sure that civil society from different localities, people from academia and technical backgrounds are present in these discussions. Use the principles to give legislators or state-actors context about the issues and good practices of content moderation.

D. Hold a press conference

showing new evidence of issues experienced on social media platforms. Use this opportunity to reference some of the agreements that companies made in your face to face meetings (when you see that they have not done enough). Tech companies do not like having a negative image in the local media. They might be keen to act swiftly to the action points, after your press conference.

E. Facilitate targeted actions towards directors of the companies:

If you do not see positive progress by the companies regarding your requests or their responses are vague, then organise social media action. Social media actions allow you to reach out to the wider public that are supporting your advocacy to send targeted messages to senior directors on social media. Many of them are connected on Twitter, Facebook, and LinkedIn and their contacts can be easily found on these platforms. Send targeted messages to them and encourage others in your networks to do the same. Or you can organise an open-letter, inviting civil society and activists to sign it with key actions you want companies to take in a set timeframe.

F. Join the Santa Clara Principles coalition in our advocacy actions.

We plan to collect more signatures and organise public meetings with tech companies. This will be an opportunity to show solidarity and have your voice be heard. The stronger the voice, the better. Follow new developments on **https://santaclaraprinciples.org**

* The terms “action” and “actioned” refer to 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.