Toolkit for Social Media Companies

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.

A Toolkit for Companies

This toolkit seeks to explain the importance of the Principles, key messages, and provide insight into how internet platforms should be implementing the Santa Clara Principles in their policies and practices.

The revised Santa Clara Principles reflect a desire amongst civil society for greater transparency, due process, cultural competency, and respect for human rights throughout the content moderation process. They are designed to obtain meaningful, public-facing transparency about all user-generated content, paid or unpaid.

This second iteration of the Santa Clara Principles is divided into Foundational and Operational Principles.

The Foundational Principles are overarching values that should be considered for all content moderation, and guide all companies in integrating human rights and due process into their policies and procedures, publishing clear and accessible rules, ensuring cultural competence when making moderation decisions, and providing transparency on state involvement in content moderation. They set out both the principle at stake and guidance as to how to implement that principle.

The Operational Principles set out specific practices for companies with respect to certain, specific stages and aspects of the content moderation process. Rather than the minimum requirements of the original Principles, the new Principles articulate general standards with respect to numbers, notice, and appeals and also include more robust requirements to be adopted as tech platforms mature, while recognizing that other platforms might not be able to meet all of the standards. Metrics of maturity include user base size, longevity in the market, technical capabilities, geographic spread, and capitalization. Additional metrics are proposed to address the special concerns raised by demands and requests from state actors, concerns that flagging processes will be abused, and the increasing role of automated processes in the identification of content and moderation actions taken.

As previously noted, many platforms endorsed or committed to abiding by the original Santa Clara Principles. However, very few platforms followed through on these commitments. While endorsement of the revised Principles is a welcome preliminary step, platforms must also fulfill their obligations to implement the Foundational Principles and should proactively maximize implementation of the revised Operational Principles as they mature. This means platforms must plan for adoption of the revised Operational Principles as they plan their growth by allocating appropriate expertise and resources equitably across, language, country, and region to expand access to due process and transparency. In doing so, platforms must ensure that considerations around human rights and due process are integrated at all stages of the content moderation process and that their rules and policies, and their enforcement, take into consideration the diversity of cultures and contexts in which their platforms and services are available and used.

To demonstrate accountability and enable oversight of their commitments going forward, platforms should publish detailed and specific operational road maps for implementation of each of the revised Principles. These road maps should include concrete goals, success metrics, and target dates. The road maps should also identify the resources a company has allocated to implementation, with specific metrics across language, country, and region. If a company believes it is unable to implement a specific principle due to a lack of maturity, it should explain why and provide an estimated time frame for when it expects to have capacity to implement the principle.

Platforms should update the public semi-annually on their progress and also describe key initiatives, discuss any changes or modifications to the road map or target deadlines, and identify both successes and roadblocks.

Platforms should identify a Santa Clara Principle liaison who, among other things, will be responsible for responding to civil society requests for information on the implementation of the revised Principles.

Key Messages and Recommendations

The revised Santa Clara Principles establish mean and aspirational standards that are designed to evoke greater transparency and accountability around how platforms design and implement their content moderation systems and the impact of such systems.

We are mindful that the revised Principles add additional obligations beyond those laid out in original Principles and that some of the revised Operational Principles may currently be unfeasible for smaller and new companies. It is not our intention to discourage competition or entrench the current resource-rich, market-dominant platforms, but instead to create a scalable framework that increases transparency and accountability for all platforms. Newer and smaller platforms should plan for compliance with the standards as they scale up.

Call to Action

Platforms must recognize the growing demand for more transparency from their users and from civil society advocates. While many platforms already engage in consultations with civil society when developing new policies or features, the feedback we have received throughout this process and beyond underscores the need for both broader consultations with a diverse range of actors from around the world and also robust implementation of the feedback platforms receive.

The revised Principles reflect the input of more than forty groups and individuals from roughly eighteen countries and several real-time group consultations conducted by partners in Africa, Latin America, India, and North America. They reflect a wide range of views on how transparency and accountability efforts by companies can be expanded to benefit a diverse range of users.

Specifically, we expect companies to demonstrate their commitment to implementing the revised Principles through concrete action, appropriate allocation of resources, and regular, detailed reporting on implementation efforts.