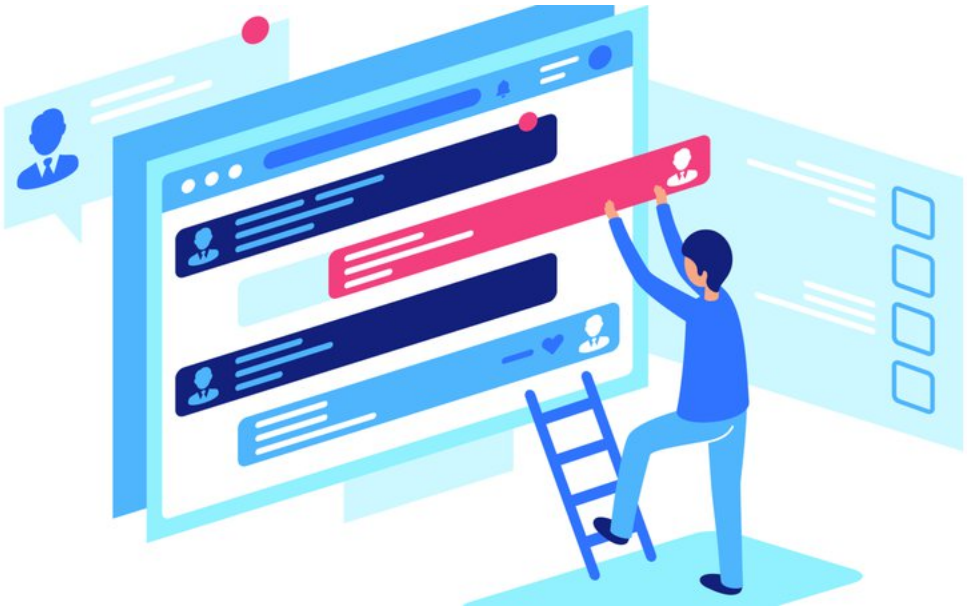
NEW AMERICA



May 2019

Assessing YouTube, Facebook and Twitter's Content Takedown Policies

How Internet Platforms Have Adopted the 2018 Santa Clara Principles

Spandana Singh

Acknowledgments

We would like to thank the drafters and signatories of the Santa Clara Principles for their support during the development of this brief.

About the Author(s)

Spandana Singh is a policy program associate in New America's Open Technology Institute.

About New America

We are dedicated to renewing America by continuing the quest to realize our nation's highest ideals, honestly confronting the challenges caused by rapid technological and social change, and seizing the opportunities those changes create.

About Open Technology Institute

OTI works at the intersection of technology and policy to ensure that every community has equitable access to digital technology and its benefits. We promote universal access to communications technologies that are both open and secure, using a multidisciplinary approach that brings together advocates, researchers, organizers, and innovators.

Contents

Introduction

One year ago, New America's Open Technology Institute, as part of a coalition of organizations, advocates, and academic experts who support the right to free expression online, released the **Santa Clara Principles on Transparency and Accountability Around Online Content Moderation**.[1] The Santa Clara Principles outline the minimum standards 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, provide meaningful due process to impacted speakers, and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights. The principles advocate for greater transparency and accountability by focusing on three key demands—comprehensive numbers detailing their content moderation activities, clear notice to affected users, and a robust appeals process.

On the first anniversary of the release of the Santa Clara Principles, we aim to assess if and how three of the largest internet platforms—YouTube, Facebook and Twitter—have implemented the recommendations outlined in the Santa Clara Principles. Our findings indicate that although the three platforms have made greater progress in implementing the recommendations related to their "notice" and "appeals" efforts, they still fall woefully short when it comes to meeting the standards set forth for the "numbers" category.

Over the past year, the online content moderation space has seen a number of important shifts that have sparked the need for such an assessment. In April 2018, Google, via YouTube, released its **first comprehensive transparency report focused on content moderation**, which provided information on YouTube's content removals based on violations of its Community Guidelines.[2] This marked the first time any internet platform published data on this aspect of its content moderation practices, and it was soon followed by similar disclosures from other internet companies.

Shortly after the release of YouTube's report, Facebook published a detailed version of its **Community Standards**, which provide an overview of the internal guidelines the company uses to make moderation decisions.[3] In addition, the company also began permitting users to appeal content takedown decisions made on individual posts, as opposed to only pages and groups. (**Here** is OTI's write-up on YouTube's initial report and Facebook's release of a detailed version of its Community Standards, with specific recommendations for how they could be improved.[4]) In May 2018, Facebook also released its first **Community Standards enforcement report** which published data on its content removals based on violations of its Community Standards.[5]

Finally, in December 2018, Twitter followed suit and began issuing **data on content removals based on violations of its Twitter Rules** in its transparency report. (**Here** is OTI's initial write-up on that report from when it first came out, with specific recommendations for how it could be improved.)

Our assessment is based on data publicly available through the platforms' transparency reports, blog posts, and press releases, as well as through responses to direct questions that we posed to the companies themselves.[6] In addition, our assessment is based on the categories of rules-violating content that companies are currently highlighting in their transparency reports and other public materials. However, we recognize that these categories do not cover the full range of rules-violating content on these platforms, and urge companies to continue expanding their transparency efforts to include all of the relevant categories.[7]

We hope that our assessment here spurs them to do an even better job providing robust transparency reports and notice and appeal processes—and that in doing so they will provide a strong example to other content platforms large and small.

Numbers

Recommendation	YouTube	Facebook	Twitter
Reports on the total number of discrete posts flagged			
Reports on the total number of discrete accounts flagged			
Reports on the total number of discrete posts removed	[1]		
Reports on the total number of discrete accounts suspended			[2]
Reports on the number of discrete posts flagged, by category of rule violated	[3]		
Reports on the number of discrete accounts flagged, by category of rule violated			
Reports on the number of discrete posts removed, by category of rule violated	[4]	[5]	
Reports on the number of discrete accounts suspended, by category of rule violated		[6]	[7]
Reports on the number of discrete posts flagged, by format or content at issue (e.g. text, audio, image, video, live stream)			
Reports on the number of discrete accounts flagged, by format or content at issue (e.g. text, audio, image, video, live stream)			
Reports on the number of discrete posts removed, by format or content at issue (e.g. text, audio, image, video, live stream)			
Reports on the number of discrete accounts suspended, by format or content at issue (e.g. text, audio, image, video, live stream)			
Reports on the number of discrete posts flagged, by source of flag (e.g. governments, trusted flaggers, users, automated etc.)	[8]		
Reports on the number of discrete accounts flagged, by source of flag (e.g. governments, trusted flaggers, users, automated etc.)			
Reports on the number of discrete posts removed, by source of flag (e.g. governments, trusted flaggers, users, automated etc.)			
Reports on the number of discrete accounts suspended, by source of flag (e.g. governments, trusted flaggers, users, automated etc.)			
Reports on the number of discrete posts flagged, by locations of flaggers and impacted users (where apparent)			
Reports on the number of accounts flagged, by locations of flaggers and impacted users (where apparent)			
Reports on the number of discrete posts removed, by locations of flaggers and impacted users (where apparent)			
Reports on the number of discrete accounts suspended, by locations of flaggers and impacted users (where apparent)			

[1] YouTube does not provide one figure for the total number of discrete posts removed. However, it does provide two separate numbers for the total number of discrete comments removed and the total number of discrete videos removed, which can be combined to ascertain the total number of discrete posts removed.
[2] Twitter reports on the number of "unique accounts actioned". The term actioned refers to a range of enforcement actions, including the suspension of accounts. However, Twitter does not break down the data for each of these enforcement options, and as a result only received partial credit.
[3] YouTube only reports on the number of discrete posts flagged, by category of rule violated, for content flagged by human flaggers, and for videos. It does not provide similar reporting for content flagged by automated tools and for comments.
[4] YouTube only reports on the number of discrete posts removed, by category of rule violated, for videos and not for comments.
[5] Facebook reports on the amount of content it took action on. This reporting covers a range of enforcement actions, including content removal. However, Facebook does not break down the data for each of these enforcement options, and as a result only received partial credit.
[6] Facebook reports on the amount of content it took action on, but does not provide a distinction in this reporting between accounts actioned and content actioned. In addition, Facebook's reporting covers a range of enforcement actions, including disabling accounts. However, Facebook does not break down the data for each of these enforcement options, and as a result only received partial credit.
[7] Twitter reports on the number of "unique accounts actioned", by category of rule violated. The term actioned refers to a range of enforcement actions, including the suspension of accounts. However, Twitter does not break down the data for each of these enforcement options, and as a result only received partial credit.
[8] YouTube currently breaks down the number of videos removed by source of flag (e.g. automated flagging, individual trusted flagger, etc.) and also provides a further break down of human flags by flagger type for videos (user, NGO, etc.). However, they, do not provide similar reporting for comments removed.

Numbers

As demonstrated by the chart above, despite the fact that YouTube, Facebook, and Twitter have all issued transparency reports that cover content takedowns based on violations of their Terms of Service, these reports do not feature many of the granular recommendations around "numbers" set forth by the Santa Clara Principles. Out of the 20 recommendations on "numbers," YouTube only fully meets four and partially meets three, Facebook fully meets none and only partially meets two, and Twitter only fully meets two and partially meets two. This demonstrates that these three platforms still have a long way to go before they are successful in providing a more granular level of transparency and accountability around their content takedown decisions.

In addition, as the chart demonstrates, there is a lack of standardization in what companies are currently reporting on in their transparency reports and there is no one recommendation with which all three companies are currently fully complying. This variation in data disclosures—including the lack of basic data around the total number of flagged and removed posts on each service—prevents meaningful comparison or combination of data across companies, which is ideally one of the key benefits of having transparency reports in the first place. We urge the companies to consider the detailed recommendations on standardization that OTI has offered in its **Transparency Reporting Toolkit** focused on content takedown reporting. [8]

Some platforms have introduced their own unique metrics, such as Facebook's prevalence metric, which measures "the estimated percentage of views that were of violating content" for each of the reported content categories. These metrics can be valuable in painting a broader picture of content takedown practices on a given platform, and in spotlighting issues related to content takedown practices that companies find important. We certainly encourage innovation in the development and implementation of such unique metrics and hope that it will continue. However, these unique metrics should supplement, not supplant, the basic standard metrics that have been set forth in the principles and described in detail in our toolkit.

Currently, YouTube's Community Guidelines enforcement report provide the most comprehensive overview of Terms of Service-related content takedown practices, as the report provides data on channel, video, and comment removals, as well as some limited data on the number of videos flagged, the role automated tools and human flaggers played in flagging and getting content removed, and the categories of rules that were violated by removed content.

Twitter's Rules Enforcement report offers the second-greatest level of transparency and granularity, as it provides data on the number of unique

accounts that were reported and the number of unique accounts that were actioned. These data points are broken down based on the category of the Twitter Rules that were violated. However, Twitter's report does not provide data related to the flagging or removal of individual pieces of content, such as tweets or comments.

Finally, Facebook's Community Standards enforcement report provides insight into the amount of content, per category of rule violated, that the company took action on. Facebook's report also provides data related to several of their own unique metrics, such as how much violating content Facebook identified before it was flagged to the company, and the previously discussed prevalence metric. However, it is overall the least successful at providing the range of basic data that the principles demand.

By issuing Terms of Service-related content takedown reports, YouTube, Facebook, and Twitter took important steps towards providing transparency and accountability around their content moderation practices. However, these reports fall woefully short when it comes to providing meaningful and granular data. In the future, these platforms should expand the data they report on, and standardize their reporting to enable meaningful comparisons and analyses.

Notice

Recommendation	YouTube	Facebook	Twitter
Provides the URL in notice		[1]	[2]
Provides content excerpt in notice			
Provides specific clause of the guidelines that the content was found to violate			
Provides how the content was detected and removed (e.g. flagged by users, government, automated detection)			
Provides explanation of the process through which the user can appeal the decision			
Notices should be available in a durable form even if a user's account is suspended or terminated		[3]	
Provides users who flag content with a log of content they have reported and the outcomes of the moderation process			

[1] Once a piece of content is suspended or removed, the URL to it ceases to function.
[2] Once a piece of content is suspended or removed, the URL to it ceases to function.
[3] If an account is suspended, a user will retain access to their account and can therefore see relevant notices. If an account is terminated, they cannot. However, if a user has opted in to email notifications, they will receive a durable form of the notice via email. Further, in some jurisdictions, such as in the European Union, Facebook is required to provide access to a full copy of information stored in a user's account if a user requests it. This would include any suspension or takedown notices.

Notice

YouTube, Facebook, and Twitter demonstrated greater progress in implementing the recommendations related to the "notice" aspect of the Santa Clara Principles.

Out of the seven recommendations put forth, YouTube fully met five of them, Facebook fully met four and partially met one, and Twitter fully met four. In the case of all three companies, the notices provided to users who have had their content removed for violating a platform's Terms of Service are relatively detailed and either fully or partially include some key pieces of information: reference to the specific rule that was violated, an explanation of how a user can appeal the takedown decision, and a durable form of the notice.

The main gap that still exists in this section of recommendations, however, is related to providing transparency around how a user's content is detected and removed. Given that automated tools are increasingly being used to identify and remove content (e.g. YouTube's Community Guidelines enforcement report highlighted that the majority of videos removed—6,190,148 in all—were detected using automated flagging tools[9]), this is an aspect of the content moderation process that companies need to provide greater transparency around.

Appeals

Recommendation	YouTube	Facebook	Twitter
Appeals are reviewed by a human or a panel of persons that were not involved in the initial decision			[1]
Appeals offer the opportunity to present additional information that will be considered in the review			
Individual appealing removal of content receives notifications of the results of the review			
Individual appealing removal of content receives reasoning sufficient to allow them to understand the decision			

[1] For some sensitive policies (e.g. child sexual exploitation, violent extremist groups), enforcement is done by specialized policy teams that also handle appeals for those difficult cases. In the scaled Twitter Rules enforcement operation, initial reviews and appeals are generally handled by different teams of reviewers. Given that some reports require specialized language skills or cultural knowledge, and within the scale that Twitter operates, this is not possible in all cases.

Appeals

YouTube, Facebook, and Twitter demonstrated the greatest level of success with the recommendations set forth in the "appeals" section of the Santa Clara Principles. Out of the four recommendations put forth in this section, YouTube fully met all four, Facebook fully met three, and Twitter fully met three and partially met one.

In the cases where Facebook failed to provide users the opportunity to present additional information during the appeals process and Twitter only partially succeeded in ensuring appeals are reviewed by a different human or panel of persons than the original moderator, both companies had a specific explanation

for these gaps, which are rooted in their specific content moderation approaches and policies. In the case of Facebook, for example, the content review process is structured so that moderators purposefully do not have access to external content (i.e. content that would not have been available to them in a traditional content moderation scenario). In the case of Twitter, initial reviews and appeals are typically handled by different teams of reviewers. However, given that some reports—such as those related to sensitive policies, language skills, or cultural knowledge—require specialized policy teams, this is not guaranteed. Overall, however, the companies have demonstrated a positive commitment to meeting the recommendations set forth in this section of the principles.

Going forward, we urge the companies to provide more information in their transparency reports around how effective their appeals processes are, and about how many users and pieces of content are impacted by these appeals processes. Facebook is already taking steps towards this, as it has committed to including in its future reporting a metric that reflects how often the company corrected its content takedown mistakes. The platform is also working to publish data on their appeals process in its transparency report as well.

Conclusion

Over the past year, YouTube, Facebook, and Twitter have taken leading steps to providing greater transparency and accountability around their content takedown practices. Although these platforms have made positive strides towards providing transparency around their notice and appeals processes, they have made minimal progress when it comes to publishing meaningful data points that illustrate the scope and scale of their content moderation efforts.

In the coming year, we hope these platforms will demonstrate a greater commitment to implementing the recommendations set forth in the Santa Clara Principles. This will not only enable the public to hold these companies accountable for their management of online speech and expression, but it will also encourage other players in the industry to similarly adopt the principles and provide transparency around their content takedown practices (**Reddit**, for example, has begun this process).

Going forward, we also hope that YouTube, Facebook, and Twitter, as three leading content regulating platforms in this space, will adopt greater standardization in the data they are reporting and the approaches they take to content takedowns. This will enable more meaningful comparisons and analyses of their efforts. In addition to this, we also welcome innovation on the part of the platforms themselves or from other organizations and individuals, in terms of the metrics and data points being reported on.

Two years ago, we knew far less than we do today about how major internet platforms are regulating and removing online content. The past year has demonstrated significant progress in shedding a light on these operations and fostering an industry-wide discussion on best practices and standards in this space. We look forward to another year of work on these issues, and hope that by 2020, the recommendations set forth in the Santa Clara Principles, and the values they are based on, are an industry-wide standard.

Notes

1 Santa Clara Principles on Transparency and Accountability Around Online Content Moderation, https://santaclaraprinciples.org/

2 YouTube, YouTube Community Guidelines enforcement, https://transparencyreport.google.com/youtube-policy/removals.

3 Facebook, "Community Standards," https://www.facebook.com/communitystandards/.

4 Kevin Bankston and Spandana Singh, "Facebook And Google Finally Take First Steps On Road To Transparency About Content Moderation," TechDirt, April 26, 2018, https://www.techdirt.com/articles/20180426/10164939724/facebook-google-finally-take-first-steps-road-to-transparency-about-content-moderation.shtml.

5 Facebook, Community Standards Enforcement Report, https://transparency.facebook.com/community-standards-enforcement.

6 Unless otherwise noted, all YouTube data in the charts was obtained from YouTube's Community Guidelines enforcement report or from direct questions posed to the company, all Facebook data in the charts was obtained from Facebook's Community Standards enforcement report, Facebook's blog post on their Community Standards, Facebook's response to an open letter penned by the Santa Clara Principles signatories or from direct questions posed to the company, and all Twitter data in the charts was obtained from Twitter's Rules enforcement report or from direct questions posed to the company.

7 YouTube's report provides data on the following categories: spam, misleading or scams, nudity or sexual content, child safety, impersonation content, content that promotes violence and violent extremism, harassment and cyberbullying, hateful or abusive content, violent or graphic content, and harmful or dangerous content. Facebook's initial report provided data on adult nudity and sexual activity, hate speech, terrorist propaganda, fake accounts, spam, violence and graphic content and since then the report has been expanded to include data on bullying and harassment and child nudity and sexual exploitation of children. Twitter's report provides data on abuse, child sexual exploitation, hateful conduct, private information, sensitive media, and violent threats.

8 Spandana Singh and Kevin Bankston, The Transparency Reporting Toolkit: Content Takedown Reporting, October 25, 2018, https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/.

9 YouTube, YouTube Community.

NEW AMERICA