6. Policing speech

Users are in the dark about the role that governments, private parties, and companies themselves play in policing the flow of information online.

Internet and mobile ecosystem companies act as powerful gatekeepers of global communication flows. Companies police content and regulate access to services according to their own private rules, and also at the request of governments and other third parties.

It is fair to expect companies to set rules prohibiting certain content or activities—like toxic speech or malicious behavior. However, when companies develop and enforce rules about what people can do and say on the internet—or whether they can access a service at all—they must do so in a way that is transparent and accountable. It is also fair to expect governments to set limits on freedom of expression for these companies to abide by, so long as those limitations are lawful, proportionate, and for a justifiable purpose, as outlined in international human rights instruments.[60] But people have a right to know how and why their speech or access to information may be restricted or otherwise shaped by companies—whether at the behest of governments, in compliance with laws, or for the companies' own commercial reasons.

The 2018 Index therefore includes six indicators measuring corporate transparency about processes for censoring online content or restricting access to their platforms or services. Collectively, these indicators evaluate company disclosure of policies and mechanisms for compliance with government requests, court orders, and other lawful third-party requests as well as for the enforcement of private rules, set by the company, about what types of speech and activities are permissible.[61] We expect companies to clearly disclose what types of content and activities they prohibit **(F3)**, and to publish data about the volume and nature of content and accounts they remove or restrict for violating these rules **(F4)**. Companies should also clearly disclose policies for responding to all types of third-party requests to restrict content and user accounts **(F5)**, and publish data about the types of such requests they receive and with which they comply **(F6, F7)**. We expect companies to notify users when they have removed content, restricted a user's account, or otherwise restricted access to content or a service **(F8)**.

6.1. Transparency and accountability ## {#section-61}

Despite some positive steps, internet and mobile ecosystem companies still don't disclose enough about their role in policing online speech.

While companies continued to make steady improvements to transparency reporting, particularly about government requests, there is still much room for improvement. Results of the 2018 Index show limited overall improvement in the past year by internet and mobile ecosystem companies in

publicly disclosing data and other information about all the ways that content is policed and managed on their platforms (Figure 14).[62]

As Figure 14 illustrates, most companies disclosed something about what content or activities are prohibited **(F3)**, while few revealed anything about actions they take to enforce these rules **(F4)**. Two companies—**Facebook** and **Tencent**—improved their disclosure of terms of service enforcement **(F3)**, but companies across the board failed to provide enough information about these practices for users to understand what actions companies take to enforce their terms of service or how these actions affect users (see Section 6.2).

Five companies—**Apple**, **Facebook**, **Telefónica**, **Twitter**, and **Oath**—improved their disclosure of how they handle government requests to censor content and restrict accounts, but all lacked key information about how they respond to such demands. (see Section 6.3).

For companies to be fully transparent with users about their role in policing content or restricting access, they must notify users in the event of content or account restrictions. They must also provide information to those who are attempting to access content that has been removed, and clearly disclose the reason why. As Figure 14 shows, companies overall lacked clear commitments to notify users when and why they remove content, with an average score of just 22 percent among internet and mobile ecosystem companies on this indicator. Three companies—**Facebook**, **Oath**, and **Twitter**—improved their disclosure of policies for notifying users when accessing content that has been removed **(F8)**. However, **Microsoft** lost points on this indicator for removing information that was previously available about policies for notifying Skype users when their accounts have been suspended.

6.2. Terms of service enforcement ## {#section-62}

Internet and mobile ecosystem companies lack transparency about what their rules are and actions they take to enforce them.

Internet and mobile ecosystem companies have come under growing pressure from policymakers and the public to better police the content that appears on their platforms due to concerns about hate speech, harassment, violent extremism, and disinformation. At the same time, companies must be transparent and accountable for how they set rules about what is allowed on their platforms and how decisions are made to enforce them. The Index contains two indicators evaluating how transparent companies are about what their rules are and how they are enforced. We expect companies to clearly disclose what types of content and activities they prohibit on their services and the process for enforcing these rules **(F3)**. We also expect companies to publish data about the volume and nature of content and accounts they have removed or restricted for violating their terms **(F4)**.

Results of the 2018 Index show that while internet and mobile ecosystem companies disclosed at least some information about what types of content or activities are prohibited by their terms of service, most disclosed nothing about the actions they took to enforce these rules (Figure 15).

Facebook disclosed more than the rest of its peers about what content and activities it prohibits and its processes for enforcing these rules **(F3)**. The company improved its disclosure of methods it uses to identify prohibited content, including user-flagging mechanisms, studying activity patterns, the use of artificial intelligence, and partnerships within industry and with civil society and governments.[63] These improvements since the 2017 Index put the company ahead of Microsoft and Kakao, which previously received the highest scores on F3.

Kakao and **Microsoft** disclosed more information than most other internet and mobile ecosystem companies, apart from Facebook. Disclosures included some information about their processes used to identify prohibited content or accounts. Both companies provided clear information about what content they prohibit and why they might restrict a user's account, as well as some information about the processes they use to identify offending content or accounts and their process for enforcing their rules.

YouTube (Google) and **Facebook** were the only social media platforms to receive full credit for their disclosure of mechanisms and processes used to identify prohibited content or activities. YouTube, like Facebook, disclosed information about a range of different types of tools it uses, including a community guidelines flagging process, staff reviews, and a system to help users identify copyrighted content.[64]

Tencent improved its disclosure by providing more examples to illustrate how it enforces its rules **(F3)**. This shows that companies operating in more restrictive environments can improve in this area without regulatory change.

Just four companies—Facebook, Google, Microsoft, and Twitter—disclosed any data about the volume or type of content or the number of accounts they restrict for violating their rules, and even these companies fell short.

Companies should not only be transparent about what the rules are but they should also reveal what actions they take to enforce them. We expect companies to clearly disclose data on the volume and nature of content or accounts they restricted for terms of service violations, including the reasons for doing so. This means reporting data on the amount of content removed for containing hate speech, pornography, or extremist content—so long as these types of content are specifically and clearly prohibited in the terms of service—as well as disclosing the number of accounts suspended and why.

Index data shows that companies are making incremental progress in this area: in the 2015 Index, no company disclosed any data about the volume or nature of content or accounts restricted for violating their rules. In the 2017 Index, three companies—**Google**, **Microsoft**, and **Twitter**—each received a small amount of credit for disclosing some data about content they removed for terms of service violations, although all still failed to provide comprehensive or systematic data on these actions.[66] In the 2018 Index, four companies—the same three companies that received credit in the 2017 Index plus **Facebook**—divulged some information about different actions they took to enforce their terms of service. But a closer look reveals serious gaps in disclosure:

Twitter: Twitter stated in a blog post that it suspended 235,000 accounts for violating its policies related to promotion of terrorism over a six-month span in 2016, but the company did not report information beyond this time period.[67] It also reported the number of times it removed content based on requests from government officials who flagged content that violated the terms of service.[68]

Microsoft: Microsoft published data about its removal of "non-consensual pornography" in breach of its terms of service, but did not report any other data about actions it took to enforce other types of terms of service violations.[69]

Google: Google gave some data on content removals from YouTube, although the data was not comprehensive or consistent. In September 2016, YouTube stated that in 2015 the company removed 92 million videos for violating its terms of service.[70] It also reported that one percent of the videos it removed were for hate speech and terrorist content.

Facebook: The company in 2017 stated that in an effort to combat the spread of misinformation, it identified and removed more than 30,000 fake accounts in France. But it did not report information about removals from any other countries or the scope of these removals in general.[71] Facebook also reported that during the months of April and May 2017, it had removed around 288,000 posts each month, globally, for containing hate speech, but it does not report this information systematically.[72]

Most internet and mobile ecosystem companies failed to disclose how they identify content or activities that violate their rules—and none revealed if they give priority to governments or other third parties to flag content or accounts that breach these rules.

While all of the internet and mobile ecosystem companies in the 2018 Index disclosed at least some information about what types of content or activities they prohibit and reasons why they might restrict a user's account, fewer disclosed clear information about what processes they use to identify offenses on their platforms. Users have a right to know whether their content might be taken down through automated processes, human reviewers, or some combination of these and other methods. Users also have a right to know whether the platforms they use give priority consideration to "flagging" by governments or private individuals.

Some companies are known to designate specific individuals or organizations for priority consideration when they report or "flag" content that violates their terms of service.[73] YouTube (Google) is credited in the 2018 Index for disclosing information about its "trusted flaggers" program, in which more robust tools are provided to "people or organizations who are particularly interested in and effective at notifying us of content that violates our Community Guidelines."[74] This program is credited in media reports with helping reduce extremist content on the platform.[75] In 2016, the European Commission announced an agreement with Facebook, Microsoft, Twitter, and YouTube (Google) to remove hate speech online, and which encourages companies to "strengthen their ongoing partnerships with civil society organisations who will help flag content."[76] In 2017, Indonesian media reported that YouTube and Twitter would allow "selected users to flag material deemed as being linked to terrorism."[77]

However, companies do not disclose much information about how these systems work in practice. While YouTube (Google) disclosed information about priority flagging processes for private parties (**F3**, Element 5), no company disclosed if they give priority flagging status to individuals employed by governments (**F3**, Element 4). Nor is it clear how or whether a company assesses the independence or motivations of a private flagger.

> **What is priority flagging?** Companies that host public or user-generated content may have systems in place to allow users to "flag" content or accounts that they think violates the company's rules. Once an item is flagged, some person (or system) at the company must decide whether to take action and if so, whether to remove or restrict access to the content, whether to take action against the user who posted it, or whether to take no action at all (for example, if the content was flagged erroneously). We expect companies to disclose information about the processes they use to identify content or activities that violate their rules, including if they use flagging mechanisms. In addition, if content or accounts flagged for violating a company's rules by a government official or a particular person or group is given extra consideration, immediate review, or prioritization through other means, we expect companies to clearly disclose this information.

Users of internet and mobile platforms have a right to know if authorities from their own government

(or any other government that they may want to criticize publicly) are availing themselves of such priority status, thereby enabling them to circumvent the process of serving the company with an official government request or court order, which would be included in company transparency reports and become a matter of public record in many countries. Information about the volume and nature of content being censored at the behest of government authorities—whether formally or informally—is essential for users to identify abuse of a platform's content policing system. Without such information it is not possible to hold companies or authorities fully and appropriately accountable when users' expression rights are violated. Yet companies keep us largely in the dark about whether governments are availing themselves, directly or indirectly, of informal flagging mechanisms.

6.3. External requests to restrict content and accounts ## {#section-63}

Companies lack transparency about how they handle formal government and private requests to censor content or restrict accounts.

Aside from platforms' private mechanisms for flagging terms of service violations, internet and mobile ecosystem companies receive a growing number of external requests to remove content or restrict user accounts via more formal and official channels. These requests come from government agencies, law enforcement, and courts, who ask companies to remove content that violates the law, infringes on someone's privacy, or contains hate speech, extremist content, or pornography. Requests can also come from self-regulatory bodies, like the UK's Internet Watch Foundation [78], or from individuals who can ask companies to remove content under the 2014 "Right to be Forgotten" ruling,[79] or through a notice-and-takedown system such as the U.S. Digital Millennium Copyright Act.[80]

> **How does RDR define government and private requests? Government requests** are defined differently by different companies and legal experts in different countries. For the purposes of the Index methodology, all requests from government ministries or agencies, law enforcement, and court orders in criminal and civil cases, are evaluated as "government requests." Government requests can include requests to remove or restrict content that violates local laws, restrict users' accounts, or to block access to entire websites or platforms. We expect companies to disclose their process for responding to these types of requests (F5), as well as data on the number and types of such requests they receive and with which they comply (F6). **Private requests** are considered, for the purposes of the Index methodology, to be requests made by any person or entity through processes that are not under direct governmental or court authority. Private requests can come from a self-regulatory body such as the Internet Watch Foundation, through agreements such as the EU's Code of Conduct on countering hate speech online, from individuals requesting to remove or de-list content under the "Right to be Forgotten" ruling, or through a notice-and-takedown system such as the U.S. Digital Millennium Copyright Act (DMCA). See Index glossary of terms at: https://rankingdigitalrights.org/2018-indicators/#Glossary

Although a handful of companies made notable improvements to their transparency reporting, as Figure 16 illustrates most companies in the 2018 Index failed to disclose sufficient information about how they handle government and private requests to censor content and restrict user access (see Section 6.1).

In general, and as was also the case in the 2017 Index, most companies tended to do better at disclosing about their *processes* for responding to government or private requests to remove content or restrict accounts **(F5)**, than they did at reporting actual *data* about the number and type of government and private requests they received and with which they complied **(F6, F7)**.

Notably, **Google** and **Facebook** earned the highest marks for disclosing their processes for responding to third-party requests, but disclosed less-comprehensive data about the number and type of requests they received **(F6-F7)**. **Apple** improved its disclosure but still failed to disclose anything about removing apps from its App Store. While Apple disclosed data on the number of requests it received from different governments to restrict or delete users' accounts, it failed to disclose any similar data about apps it removed from its app store, or the subject matter associated with these removals **(F7)**. According to reports, Apple has removed apps from its App Store in China, Russia, and elsewhere—including the apps for *The New York Times* and LinkedIn,[81] Skype,[82] and hundreds of VPNs—in response to requests from governments.[83]

There were also notable blind spots around companies' handling of private requests. Companies tended to report less information about the number of private requests they received to remove content **(F7)** compared to those they received from governments **(F6)**. This means users have less information about whether and under what circumstances companies are complying with private requests to censor content or restrict user accounts, or the volume of these types of requests that companies receive.

However, **Twitter, Kakao, Microsoft**, and **Yandex** disclosed more data on private requests than on government requests:

Twitter, for example, disclosed data about the copyright and trademark takedown requests it received, and the number of removals as part of the "EU Trusted Reporters" program to comply with local hate speech laws in Europe. It disclosed the reasons associated with these requests and the number of requests with which it complied.

Microsoft disclosed data on requests to remove information from the Bing search engine, in line with the "Right to Be Forgotten" ruling, as well as removal requests due to alleged copyright infringement. For both of these types of requests, Microsoft disclosed the number of URLs for which it received takedown requests and with which it complied.

Kakao provided data about several different types of private requests, including requests to remove content due to copyright or trademark violations, or defamation. Kakao also listed the number of requests with which it complied.

6.4 Recommendations for companies ## {#section-64}

Publish transparency reports that include comprehensive data about the circumstances under which content or accounts may be restricted. Transparency reports should ideally be published every six months. Information should include:

Government requests to restrict content or accounts: In particular, companies should disclose the number of requests they receive per country as well as the number of requests with which they comply.

Private requests to restrict content or accounts: Companies should disclose the volume and

nature of requests received, and number complied with, from private individuals or entities not connected to official government or court processes. Companies should also disclose information about the circumstances under which they will respond to private requests, and that they conduct due diligence on such requests.

Priority flagging: If any organizations or individuals are given special consideration when flagging content for removal as part of informal private processes that do not involve lawful government requests or court orders, these entities should be listed, or at least a description of the process for designating "priority flaggers" should be disclosed. Numbers of requests received from different types of priority flaggers should also be reported, with as much granularity as possible. If a company does not receive or entertain a particular type of request, or if it doesn't entertain requests from certain types of third parties (e.g., private individuals acting without legal authority), the company should also clearly disclose that information.

Terms of service enforcement: Companies should disclose the number of actions taken to remove content or restrict accounts that violated the company's rules, and the reasons for doing so (e.g. the number of accounts restricted for posting extremist content, the number of items removed for containing hate speech, etc.).

Provide examples of how rules are enforced. Even when companies publish their rules, it is very unclear how they are enforced. Reports of arbitrary blocking or inconsistent restrictions on accounts make it all the more difficult to understand how platforms are being policed. Clearer disclosure on this front will help restore trust between users and the services on which they rely, and could help empower users to understand and seek remedy when their content or account has been unfairly restricted.

Commit to notify users of censorship events. Companies should disclose their policies for notifying users when they restrict content or accounts, including the reason for doing so.

Footnotes

[60] "Universal Declaration of Human Rights" (United Nations, December 10, 1948), http://www.un.org/en/universal-declaration-human-rights/ and "International Covenant on Civil and Political Rights" (United Nations, December 16, 1966), http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx/.

[61] "2018 Indicators: Freedom of Expression," see F3-F10: https://rankingdigitalrights.org/2018-indicators/#F.

[62] See Chapter 5, 2017 Corporate Accountability Index, Ranking Digital Rights, https://rankingdigitalrights.org/index2017/assets/static/download/RDRindex2017report.pdf.

[63] "Hard Questions: How We Counter Terrorism," Facebook Newsroom, June 15, 2017, https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/.

[64] YouTube, "How Content ID Works," accessed March 19, 2018, https://support.google.com/youtube/answer/2797370/.

[65] 2015 Corporate Accountability Index, Ranking Digital Rights, p. 25, https://rankingdigitalrights.org/index2015/assets/static/download/RDRindex2015report.pdf.

[66] See p. 29-30, 2017 Corporate Accountability Index, Ranking Digital Rights, https://rankingdigitalrights.org/index2017/assets/static/download/RDRindex2017report.pdf.

[67] “An update on our efforts to combat violent extremism,” Twitter Blog, August 18, 2016, https://blog.twitter.com/official/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html.

[68] Twitter Government TOS Report, https://transparency.twitter.com/en/gov-tos-reports.html.

[69] Microsoft Content Removal Requests report, https://www.microsoft.com/en-us/about/corporate-responsibility/crrr/.

[70] “Why flagging matters,” YouTube Official Blog, September 15, 2016, https://youtube.googleblog.com/2016/09/why-flagging-matters.html.

[71] “Improvements in protecting the integrity of activity on Facebook,” Facebook Security, April 12, 2017. https://www.facebook.com/notes/facebook-security/improvements-in-protecting-the-integrity-of-activity-on-facebook/10154323366590766.

[72] Richard Allan, “Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?” Facebook Newsroom, June 27, 2018, https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/.

[73] “Growing Our Trusted Flagger Program into YouTube Heroes,” Official YouTube Blog, September 22, 2016, https://youtube.googleblog.com/2016/09/growing-our-trusted-flagger-program.html.

[74] “Growing Our Trusted Flagger Program into YouTube Heroes,” Official YouTube Blog, September 22, 2016, https://youtube.googleblog.com/2016/09/growing-our-trusted-flagger-program.html.

[75] Press Association, “YouTube Introduces New Measures to Curb Extremist Video Online,” *The Guardian*, June 19, 2017, https://www.theguardian.com/technology/2017/jun/18/more-must-be-done-about-extremist-content-online-says-google/.

[76] “Press Release - European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech,” European Commission, May 31, 2016, http://europa.eu/rapid/press-release_IP-16-1937_en.htm.

[77] Ismira Lutfia Tisnadibrata, “Indonesia: Google, Twitter Agree to Tighten Content Monitoring,” *BenarNews*, August 4, 2017, https://www.benarnews.org/english/news/indonesian/indonesia-terrorism-08042017183754.html.

[78] “What We Do,” Internet Watch Foundation, accessed March 22, 2018, https://www.iwf.org.uk/what-we-do.

[79] “Fact sheet on the Right to be Forgotten Ruling,” European Commission, https://www.inforights.im/media/1186/cl_eu_commission_factsheet_right_to_be-forgotten.pdf.

[80] For more information on notice-and-takedown, as well as the DMCA, see Rebecca MacKinnon et al., “Fostering Freedom Online: The Role of Internet Intermediaries” (UNESCO, 2014), http://unesdoc.unesco.org/images/0023/002311/231162e.pdf.

[81] Farhad Manjoo, “Clearing Out the App Stores: Government Censorship Made Easier,” *The New York Times*, January 18, 2018, https://www.nytimes.com/2017/01/18/technology/clearing-out-the-app-stores-government-censorship-made-easier.html.

[82] Paul Mozur, “Skype Vanishes From App Stores in China, Including Apple’s,” *The New York Times*, November 21, 2017, https://www.nytimes.com/2017/11/21/business/skype-app-china.html.

[83] Tim Bradshaw, “Apple drops hundreds of VPN apps at Beijing’s request,” *Financial Times*, November 21, 2017, https://www.ft.com/content/ad42e536-cf36-11e7-b781-794ce08b24dc/.