Submission to UN Special Rapporteur for Freedom of Expression and Opinion David Kaye: Content Regulation in the Digital Age

Ranking Digital Rights

December 15, 2017

by

Ilana Ullman, Policy and Communications Analyst

Laura Reed, Senior Research Analyst and Coordinator

Rebecca MacKinnon, Director

Table of Contents

Executive Summary 2

About Ranking Digital Rights 3

Introduction 4

1. Transparency 6

2. Impact assessment 12

3. Grievance and Remedy 15

Conclusion 18

Recommendations 19

Appendix: Full text of relevant 2017 Corporate Accountability Index indicators 21

Executive Summary

Internet platforms around the world are under increasing pressure from governments and many other stakeholders to remove objectionable content. From misinformation to hate speech to extremist content to harassment and abuse, these platforms have become gatekeepers of what information the public can access and share. Yet, despite the important role they play in mediating public discourse, and despite progress by some companies in recent years in disclosing policies and actions related to government requests, the process of policing content on internet platforms remains unacceptably opaque. As a result, users of internet platforms cannot adequately understand how their online information environment is being governed and shaped, by whom, under what authorities, for what reason. When transparency around the policing of online speech is inadequate, people do not know who to hold accountable when infringements of their expression rights occur. This situation is exacerbated by the fact that some of the world's most powerful internet platforms do not conduct systematic impact assessments of how their terms of service policies and enforcement mechanisms affect users' rights. Furthermore, grievance and remedy mechanisms for users to report and obtain redress when their expression rights are infringed are woefully inadequate.

In this submission to the UN Special Rapporteur on Freedom on Expression and Opinion David Kaye for his upcoming report on content regulation in the digital age, Ranking Digital Rights proposes the following recommendations for companies and governments:

1. **Increase transparency of how laws governing online content are enforced via internet intermediaries and how decisions to restrict content are being made and carried out.** Companies should disclose policies for decision making around content restriction, whether restriction is requested by governments, private actors, or carried out at the company's own initiative to enforce its terms of service. They should also disclose data on the volume and nature of content being restricted or removed for the full range of reasons that result in restriction. Governments must encourage if not require such transparency and match it with transparency of their own regarding demands – direct as well as indirect – that they place upon companies to restrict content.

2. **Broaden impact assessment and human rights due diligence in relation to the regulation and private policing of content.** Companies must conduct human rights impact assessments that examine policies and mechanisms for identifying and restricting content, including terms of service enforcement and private flagging mechanisms. They must disclose how such assessments are used identify and mitigate any negative impact on freedom of expression that may be caused by these policies and mechanisms. Governments should also assess existing and proposed laws regulating

content on internet platforms to ensure that they do not result in increased infringement of users' freedom of expression rights.

3. **Establish and support effective grievance and remedy mechanisms to address infringements of internet users' freedom of expression rights.** When content is erroneously removed or a law or policy is misinterpreted in a manner that results in the censorship of speech that should be protected under international human rights law, effective grievance and remedy mechanisms are essential to mitigating harm. Adequate mechanisms are presently lacking on the world's largest and most powerful internet platforms. Governments seeking increased policing of extremist and violent content by platforms should not only support but participate in the development of effective grievance and remedy mechanisms.

About Ranking Digital Rights

Ranking Digital Rights (RDR) is a non-profit research initiative housed at New America's Open Technology Institute, working with an international network of partners to set global standards for how companies in the information and communications technology (ICT) sector should respect freedom of expression and privacy. For more about the project see: https://rankingdigitalrights.org

In 2015, RDR launched its inaugural Corporate Accountability Index which evaluated 16 companies based on 31 indicators focused on corporate disclosure of policies and practices that affect users' freedom of expression and privacy. In March 2017, Ranking Digital Rights released the second edition of its Corporate Accountability Index, which ranked 22 companies according to an expanded list of 35 indicators.
The full Index results, including the report and raw data for researchers to download and use, can be found at: https://rankingdigitalrights.org/index2017/

The 2017 Corporate Accountability Index evaluated 22 internet, mobile, and telecommunications companies on 35 indicators assessing companies' public disclosures and commitments in three categories: governance, freedom of expression, and privacy. Of the 22 companies evaluated, ten provide search and/or social platforms (all services for which the companies were evaluated, including non-platform services, are listed below):

- **Baidu (China)** — Baidu Search, Baidu Cloud, Baidu PostBar
- **Facebook (US)** — Facebook, Instagram, WhatsApp, Messenger
- **Mail.Ru (Russia)** — VKontakte, Mail.Ru email, Mail.Ru Agent

- **Microsoft (US)** — Bing, Outlook.com, Skype
- **Kakao (South Korea)** — Daum Search, Daum Mail, KakaoTalk
- **Google (US)** — Search, Gmail, YouTube, Android mobile ecosystem
- **Tencent (China)** — QZone, QQ, WeChat
- **Twitter (US)** — Twitter, Periscope, Vine
- **Yahoo (US)** — Yahoo Mail, Flickr, Tumblr
- **Yandex (Russia)** — Yandex Mail, Yandex Search, Yandex Disk (cloud storage)

The full methodology, along with research guidance, can be found here: https://rankingdigitalrights.org/2017-indicators/ This submission covers results from five indicators (F3-F7) that evaluates corporate transparency about a range of company actions that affect users' freedom of expression, plus two indicators that evaluate different aspects of governance: G4 evaluates disclosure about impact assessments, and G6 evaluates disclosure of grievance and remedy mechanisms.

The third Corporate Accountability Index, which covers the same 22 companies with the same methodology, will be released in April 2018. While research for the 2018 Index is already underway, it is far from final as of the December 2017 deadline for this submission. The information in this submission is therefore based on the 2017 Corporate Accountability Index, published in March 2017.

Introduction

Companies are subject to government regulation and also self-regulate. Self-regulation is sometimes carried out unilaterally by a company, and is sometimes carried out in conjunction with other stakeholders with government support as an alternative to direct regulation (sometimes referred to as "co-regulation").[1] Of the 35 indicators used to evaluate companies in the 2017 Ranking Digital Rights Corporate Accountability Index, ten are directly relevant to the question of how content regulation affects freedom of expression on internet platforms. The results of the 2017 Index show that companies disclose the greatest amount of information about their commitments, policies, and processes in response to direct *government* demands to remove or restrict content or deactivate user accounts. They disclose the least amount of information about how *private* rules and mechanisms for self- and co-regulation are formulated and carried out. The results of RDR's 2017 Index also show that while many companies in the Index conduct assessments of how government demands affect their users' rights, few companies appear to conduct impact assessments to identify how their own private enforcement policies and practices affect the freedom of expression of users around the world.

[1] See p. 54, http://unesdoc.unesco.org/images/0023/002311/231162e.pdf

RDR's evaluation also reveals that grievance and remedy mechanisms offered by companies to users are far from adequate.

Concerns that global internet platforms are censoring speech that should be protected under international human rights law in their efforts to comply with growing government regulatory pressures are reflected in recent debates and critiques of Germany's "NetzDG" law, which came into force in October 2017.[2] The law imposes significant fines on social media platforms that do not remove hate speech content, which is illegal under German law, within 24 hours. Companies that repeatedly fail to delete this illegal content in a timely manner could face fines of up to €50 million.[3] The full impact of the law remains to be seen, but is even more important to understand in the context of other governments, such as France and the UK considering new fines for social media sites that do not remove extremist content, and increasing public pressure in the U.S. to clamp down on white supremacist content.[4] Hasty and blunt application of enforcement mechanisms being used in response to such laws has resulted in restriction of protected speech. One well-known example is the recent deletion of Syrian opposition videos on Youtube.[5] Such cases highlight why greater transparency, impact assessment and remedy are vital as governments and companies struggle with the question of how to protect the public and users from hate speech and extremist attacks without violating internet users' expression rights.

As companies strengthen policies and beef up technical mechanisms to police such content, lack of transparency about these mechanisms, or about the volume and nature of content being removed by them, makes it impossible for stakeholders to know whether these policies and mechanisms are achieving their intended purposes. Lack of impact assessment means that companies may not themselves have a clear understanding about the potential collateral damage their policies and mechanisms may inflict and how to mitigate negative impacts on users' freedom of expression. Lack of adequate grievance and remedy mechanisms means that when activists or journalists or others exercising their free expression rights are silenced by a platform using mechanisms intended to silence hate speech and extremism, they do not have reliable recourse to a process for having their case reconsidered and content reinstated.[6]

---

[2] https://edri.org/eu-action-needed-german-netzdg-draft-threatens-freedomofexpression/

[3] https://qz.com/1090825/germanys-new-social-media-law-analysis-facebook-twitter-youtube-to-remove-hate-speech-in-24-hours-or-face-fines/

[4] https://www.theverge.com/2017/6/13/15790034/france-uk-social-media-fine-terrorism-may-macron and http://thehill.com/policy/technology/347173-tech-companies-crack-down-on-hate-speech-after-charlottesville

[5] https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html

[6] https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/

1. Transparency

Findings from RDR's 2017 Index highlight specific areas of weakness as well as some emerging practices. Eight indicators in the Index evaluate company disclosures about their policies and mechanisms related to government requests and legal compliance as well as the enforcement of private rules, set by the company, about what types of speech and activity are permissible.[7] The 2017 Index results spotlight key areas in which internet platforms can improve transparency about their content moderation policies and practices. We found that companies overall lack transparency about their policies affecting users' freedom of expression—and in particular about what types of content are prohibited and what their process is for enforcing these rules. Companies also tended to disclose more information about requests they receive from governments and private parties to restrict or delete content or deactivate accounts than about actions companies themselves took to enforce terms of service. Companies also lacked disclosure of whether they notify users when they restrict content or accounts.

1.1 Transparency about company actions that affect freedom of expression lags behind transparency about actions that affect privacy. Transparency reports are one way for companies to regularly publish data about third party requests they receive and comply with, for both content removals and user information, and are increasingly a common practice.[8] Although more companies have begun issuing transparency reports, companies tend to report more information about requests they receive to share user data (affecting privacy) than they do for actions that affect freedom of expression, such as content restriction and removal or account deactivation.

Figure 1 on the next page compares company scores on transparency reporting about requests that they receive to share user information (dark blue)[9] versus transparency reporting about requests they receive to restrict or remove content, or deactivate accounts (light blue)[10].

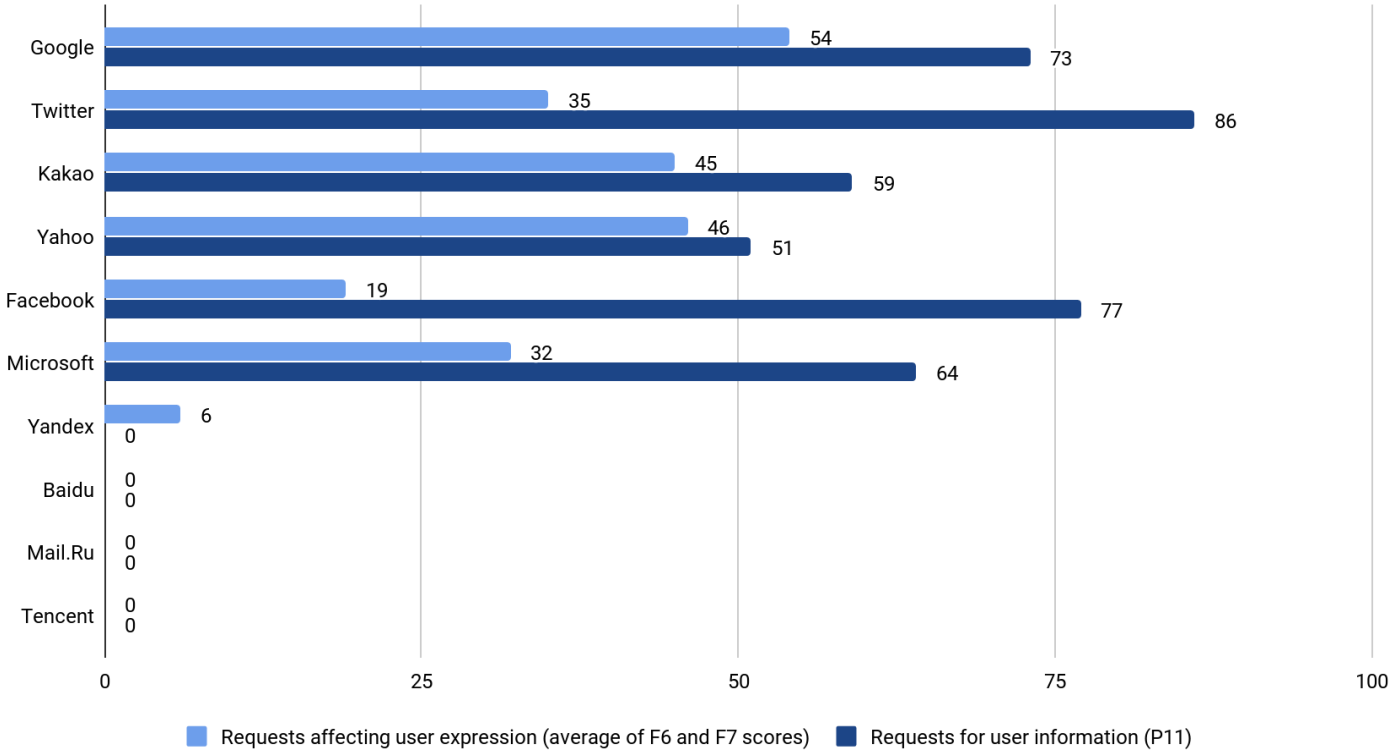
[7] See F1-F8 at https://rankingdigitalrights.org/2017-indicators/#F

[8] https://www.newamerica.org/in-depth/getting-internet-companies-do-right-thing/case-study-3-transparency-reporting/

[9] See https://rankingdigitalrights.org/index2017/indicators/#P11

[10] See https://rankingdigitalrights.org/index2017/indicators/#F6 and https://rankingdigitalrights.org/index2017/indicators/#F7

Figure 1: Transparency about government and private requests affecting user expression and privacy

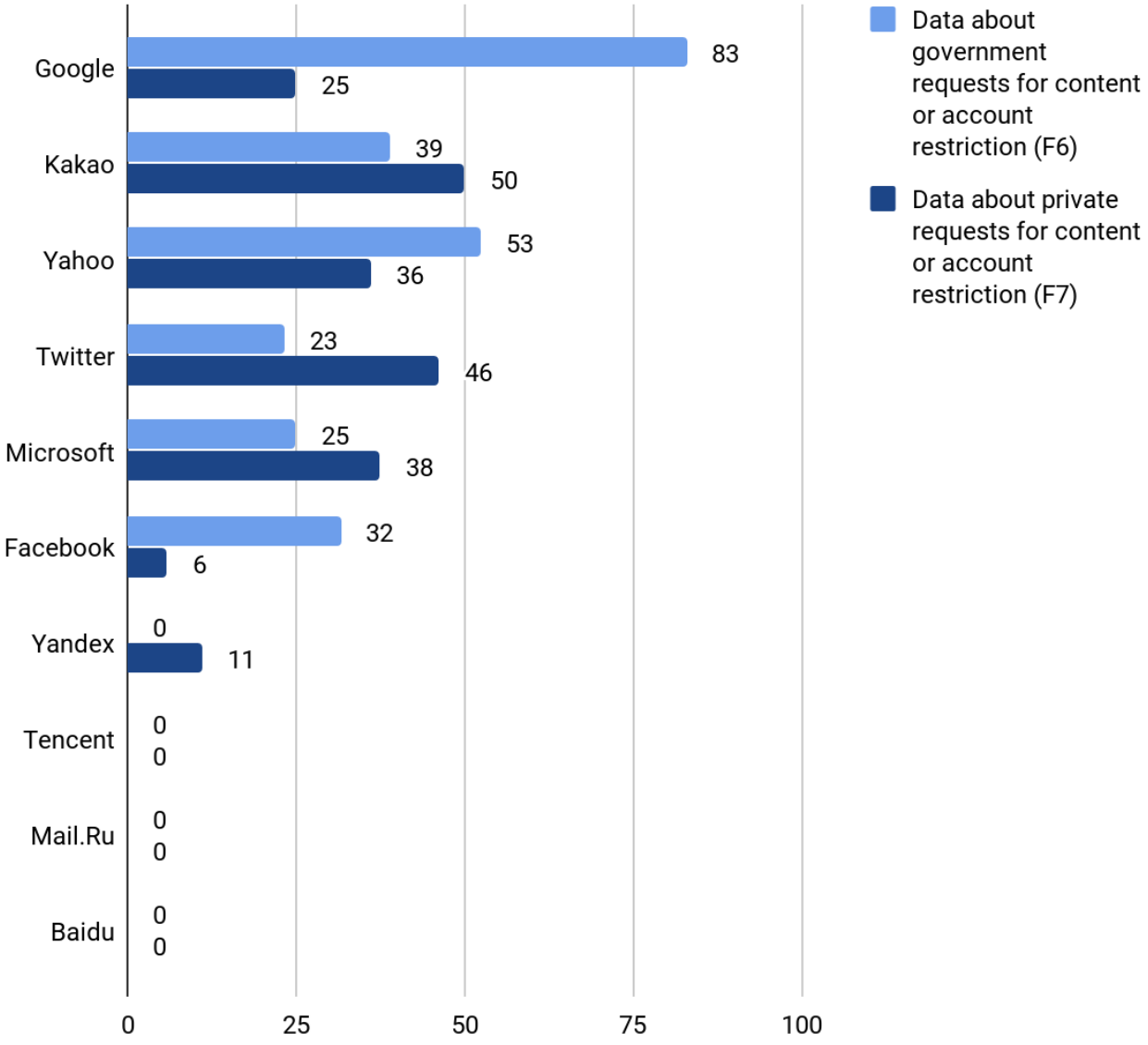


1.2 Companies vary widely in their disclosures about government requests versus disclosures about requests from private parties, including requests related to terms of service violations. All platform companies have room to improve their transparency about requests affecting freedom of expression, and should be just as transparent about government requests to remove content as they are with private ones (and vice versa). Currently, users are left with an incomplete picture of the scope and potential impact that government and private requests may have on their speech.

Note that RDR defines "government requests" for content removals as requests originating from government ministries or agencies, law enforcement, or court orders in criminal and civil cases. RDR defines "private requests" as those made by any person or entity not acting under direct governmental or court authority. Examples of private requests include requests from a self-regulatory body such as the UK's Internet Watch Foundation, or a notice-and-takedown system, such as the U.S. Digital Millennium Copyright Act.

Figure 2 below compares company scores on data disclosed about government requests (F6)[11] with their scores on data disclosed about private requests (F7).[12]

Figure 2: Data about government requests and private requests affecting user expression



Disclosure about private requests for content restriction is also important for monitoring the full impact of government requests for content restriction, given that governments often

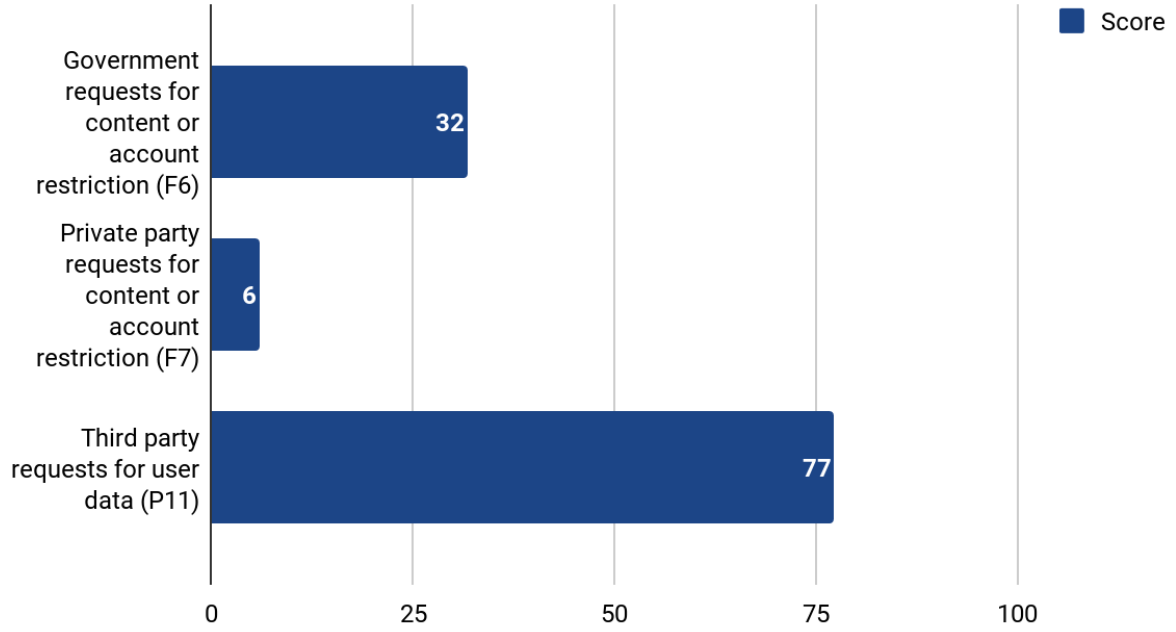
[11] See https://rankingdigitalrights.org/index2017/indicators/#F6
[12] See https://rankingdigitalrights.org/index2017/indicators/#F7

delegate take-down requests and the reporting of terms-of-service violations to private parties. For example, there have been documented cases of copyright enforcement mechanisms being abused by governments, such as Ecuador President Rafael Correa, who used millions of dollars of public funds to hire a foreign company to help delete information critical of him from sites including **YouTube**, **Facebook**, Vimeo, and Dailymotion.[13]

Of particular note is **Facebook**'s minimal disclosure of data related to any type of private requests for content restriction (F7). Its score on this indicator, 6% out of a possible 100%, reflected the least amount of disclosure in comparison with its peers, except for three companies headquartered in much more repressive speech environments (China and Russia).

Figure 3 below compares Facebook's relatively strong performance in disclosing data about government requests for user data, in contrast to its poor performance in disclosing company actions that affect users' freedom of expression.

Figure 3: Facebook transparency reporting



All platform companies have room to improve their transparency about requests they receive and how those requests are handled. They should be just as transparent about government requests to remove content as they are with private ones (and vice versa). Currently, users are left with an incomplete picture of the scope and potential impact that government and private

[13] https://www.buzzfeed.com/jamesball/ecuadors-president-used-millions-of-dollars-of-public-funds

requests to remove content may have on their speech, and more broadly on the global flow of information online.

1.3 ***Companies disclose more information about processes than data about the volume and nature of restrictions and removals.*** There continues to be a gap in what companies disclose about their role in censoring online content, with companies disclosing more about their *processes* for responding to government and private requests to restrict content and accounts than they do about the number of requests they receive and with which they comply. This makes it unclear how these processes are applied in practice, and difficult to determine the scope and potential impact of these content restrictions on freedom of expression. In light of the recent increase in government pressure to remove extremist, hate speech, and other objectionable categories of content, it is particularly important that companies publish data on the content restriction requests they receive and comply with, so that advocates and the public can determine if content removal requests are on the rise, and whether or not the company is pushing back against them.

In addition, when governments boast of increased cooperation with platform companies, this data may also offer some insight to help determine whether these claims have merit, or are merely more rhetoric. For example, in April 2017, the Vietnamese government reported it had reached an agreement with **Facebook** for it to censor content that violates local laws or posts with "fake content" about government officials.[14] Vietnamese officials claimed that Facebook had agreed to set up a direct channel of communication with the government to facilitate these requests, although Facebook stated that its process for receiving and responding to government requests is consistent across jurisdictions.[15]

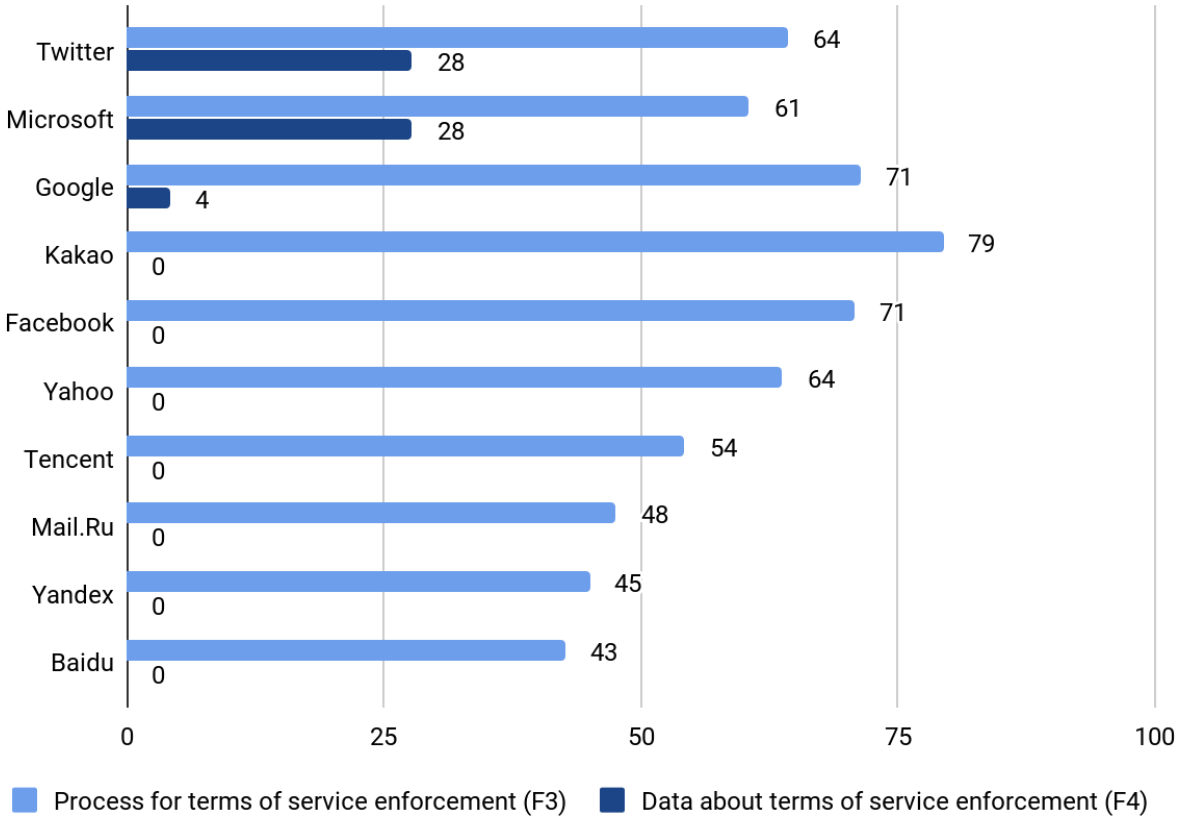
Notably, in its Government Requests Report, Facebook discloses only the number of pieces of content it restricted due to government requests, per country.[16] It does not disclose the number of requests it received, making it impossible to tell to what degree the company is complying with, or perhaps pushing back against, government requests it receives to censor content on its platform.

This gap between disclosure of content policies and how they are enforced in practice is particularly noticeable in the case of terms of service enforcement, for which the difference in company performance among the two indicators is particularly pronounced.

[14] https://www.reuters.com/article/us-facebook-vietnam/vietnam-says-facebook-commits-to-preventing-offensive-content-idUSKBN17T0A0
[15] https://www.reuters.com/article/us-facebook-vietnam/vietnam-says-facebook-commits-to-preventing-offensive-content-idUSKBN17T0A0
[16] https://govtrequests.facebook.com/

As Figure 4 below illustrates, while all companies disclose information about processes, only three companies were found to have disclosed any data about the volume and nature of content they removed at their own initiative when enforcing their terms of service:

Figure 4: Disclosure of terms of service enforcement



Of the three companies—**Google**, **Microsoft**, and **Twitter**—that disclosed any data about content they restricted as a result of enforcing their terms of service,[17] disclosure was limited to specific areas:

- In a February 2016 blog post, **Twitter** disclosed that "Since the middle of 2015 alone, we've suspended over 125,000 accounts for threatening or promoting terrorist acts,"[18] and in a follow-up post six months later it announced that "we have suspended an additional 235,000 accounts."[19] In March 2017 (after the cutoff date for information evaluated in the 2017 Index) Twitter began including terms of service-related takedowns in its transparency reports for requests that originated from governments

[17] See https://rankingdigitalrights.org/index2017/indicators/#F4
[18] https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html
[19] https://blog.twitter.com/official/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html

and concerned extremist content. In September 2017, Twitter expanded this category to include three additional terms of service violation categories, such as abusive behavior, copyright, and trademark. However, this data is limited to requests that originated from governments and does not include requests from non-governmental parties. As Twitter has recently announced upcoming changes to its rules,[20] specifically those relating to abuse, it is important that it continue to publish this data, and broaden the scope to include non-government requests, to help users and advocates evaluate the new rules' impact.

- **Microsoft** published terms of service enforcement data for "revenge porn" content (this content is illegal in some jurisdictions but not all), and the company specifically states that they remove reported links to photos and videos from search results on **Bing** "...when we are notified by an identifiable victim" (which would indicate that these are non-government requests). However, it did not publish data relating to other types of content or activities in the company's terms of service.

- In a September 2016 blog post,[21] **YouTube** disclosed that it had removed 92 million videos for violating its terms of service and that 1% of the videos removed were for hate speech and terrorist content. However, it did not provide exact numbers, and this data is not reported in an ongoing manner.

Though disclosure on this indicator is incredibly low, it still marks an improvement compared to the 2015 Index,[22] in which no company evaluated disclosed any data about content that was restricted from enforcing their terms of service.

2. Impact assessment

Since its inception in 2015, the RDR Corporate Accountability Index has consistently found that while some of the world's most powerful internet platforms publicly disclose that they carry out impact assessments on how compliance with government laws and policies may affect the freedom of expression rights of users, companies disclose little about assessing the risks to freedom of expression posed by the enforcement of their own policies.

Significant challenges to freedom of expression and privacy can arise when a company decides to introduce a new feature, launch a new service, or enter a new market. One indication that a

[20] https://blog.twitter.com/official/en_us/topics/company/2017/safetycalendar.html
[21] https://youtube.googleblog.com/2016/09/why-flagging-matters.html
[22] https://rankingdigitalrights.org/index2015/

company is considering the potential human rights implications of its policies and services is whether it discloses that it conducts human rights impact assessments (HRIAs). The UN Guiding Principles on Business and Human Rights, which articulate businesses' responsibility to respect human rights, specifically spells out companies' obligation to assess actual and potential human rights impacts and to act upon the findings.[23] Human rights impact assessments provide companies with a means to identify areas of concern in order to mitigate or prevent potential infringements on human rights, or to provide remedy for violations that may have already occurred.[24]

Many of the issues at the intersection of human rights and technology making headlines today can be traced back to a company's failure to anticipate the negative implications of its business decisions. For example, a ProPublica investigation of **Facebook**'s internal content moderation policies revealed confusing rules that protected categories such as "white men" but not "black children." Another rule, which the company said was no longer in effect, prohibited content supporting "violence to resist occupation of an internationally recognized state."[25] ProPublica reported several instances of journalists and activists in Palestine, Kashmir, Crimea, and Western Sahara, who had their content or accounts restricted as a result of this policy. Such problems raise questions about whether the company carried out any sort of impact assessment before enacting these rules and related enforcement processes. Indeed, Facebook's score on Index indicator G4, which examines company disclosures on impact assessment, shows no disclosure about impact assessment on terms of service policy formulation or enforcement.[26]

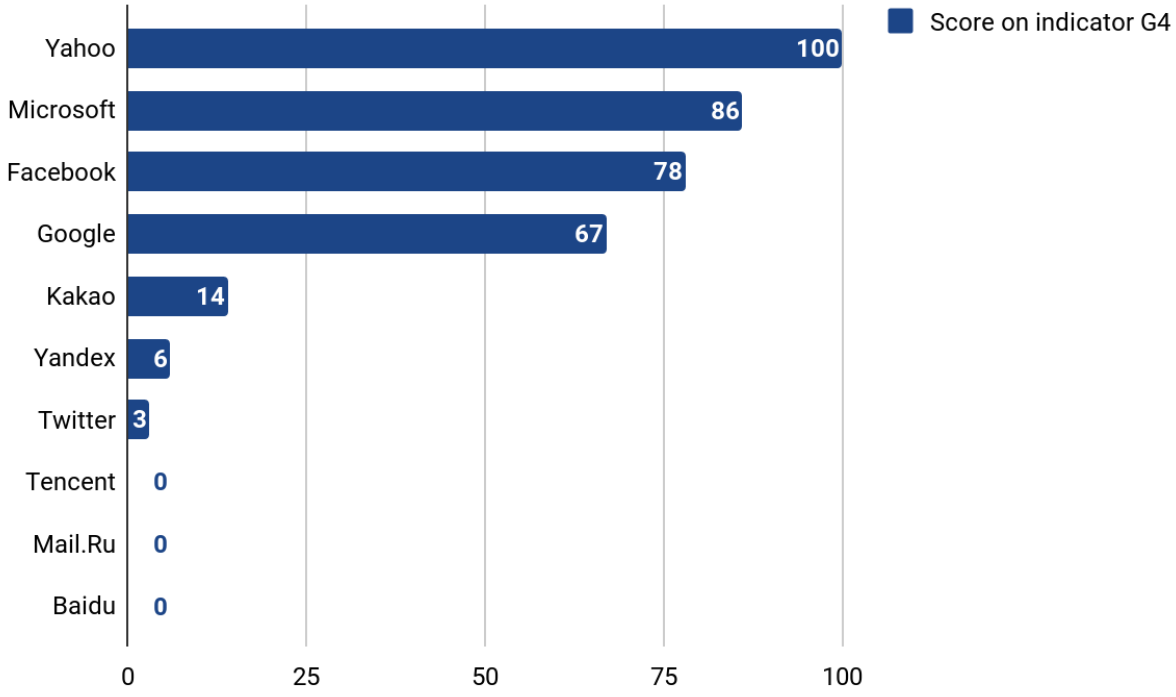
When crafting rules about what types of speech is forbidden from their platforms and in what context, companies should consult with external stakeholders and carefully examine if these rules protect those most likely to be discriminated against, and determine the impact on freedom of expression of their enforcement. It is important for companies to carry out human rights impact assessments on a regular basis and continue to assess the potential impact of their products and services after the initial rollout, and to conduct meaningful outreach to different stakeholder groups in order to learn how these technologies are surfacing varying challenges for a range of user groups. As noted in the call for submissions, the standards and processes that a platform company uses to enforces its own rules—such as those articulated in

[23] Principle 17, UN Guiding Principles on Business and Human Rights, p. 17: http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.
[24] UN Guiding Principles on Business and Human Rights, p. 18: http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf.
[25] https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms
[26] See https://rankingdigitalrights.org/index2017/indicators/#G4 ; https://rankingdigitalrights.org/2017-indicators/#G4 and https://rankingdigitalrights.org/index2017/companies/facebook/

terms of service or community guidelines policies—have significant implications for human rights. Recognizing this risk, companies should pay close attention to the implications of their terms of service enforcement policies when conducting HRIAs.

Figure 5 below illustrates the extent to which companies conduct thorough due diligence across the full range of risks to users' freedom of expression:

Figure 5: Impact Assessment



Seven of the ten platform companies evaluated in the 2017 Index disclosed at least some information about whether they carry out regular, comprehensive, and credible due diligence, for example, in the shape of human rights impact assessments (HRIA), to examine how all aspects of their business affect freedom of expression and privacy and to mitigate any risks posed by those impacts.[27] Four fully disclosed that as part of their decision-making, they consider how laws affect freedom of expression in jurisdictions where they operate. Five companies received credit for at least some information about assessing freedom of expression and privacy risks associated with existing products and services, and six disclosed at least some information about assessing freedom of expression and privacy risks associated with a new activity. However, of the ten platform companies, only two—**Yahoo** and **Microsoft**—disclosed any information about assessing and mitigating the freedom of expression and privacy risks associated with the processes and mechanisms used to enforce their terms of service.

[27] https://rankingdigitalrights.org/index2017/indicators/#G4

- **Yahoo** (now part of a new company, Oath, owned by Verizon), received full credit for disclosing that one of the circumstances triggering an HRIA is the "review of internal processes or mechanisms to enforce policies, such as our terms of service, that may impact users' rights to privacy or free expression."[28]

- **Microsoft** received partial credit for disclosure about efforts the company has taken to address the freedom of expression impacts of how it enforces its terms of use with regard to terrorist content reports. In its disclosure, Microsoft also states that its "decisions and actions in the enforcement of the terms of use for our services do not change based on whether the referral is made by a government or any other non-government entity or person."[29] However, this is only one example and is not a clear commitment that this kind of assessment occurs for all its products and services on an ongoing basis.

With the exceptions noted above, social and search platform companies are not disclosing if they conduct risk assessments of the freedom of expression risks associated with enforcing their terms of service. Much of company disclosure tends to focus on their processes for responding to government requests to censor content—companies are far less transparent when it comes to the impact that their own rules, and private party requests to censor content, can have on freedom of expression. If companies are already conducting such assessments in these areas, they should better communicate them to their users. Internet platforms must be making informed decisions about their content policies—unless they consult with affected stakeholders and thoroughly consider and take efforts to mitigate human rights harms, these policies may have significant negative consequences, which in many cases could have otherwise been anticipated.

3. Grievance and Remedy

Companies should provide meaningful remedy when they become aware of an instance in which their business operations may have resulted in the infringement of users' rights, yet few companies disclose much information about their remedy mechanisms, if they have them at all. The UN Guiding Principles on Business and Human Rights state that when a company identifies a situation in which it has caused or contributed to adverse impacts on human rights, "[...] its responsibility to respect human rights requires active engagement in remediation, by itself or in

[28] https://yahoobhrp.tumblr.com/post/75507678786/human-rights-impact-assessments-yahoo-has
[29] Microsoft Salient Human Rights Issue Report http://download.microsoft.com/download/0/0/6/00604579-134B-4D0E-97C3-D525DFB7890A/Microsoft_Salient_Human_Rights_Issues_Report-FY17.pdf

cooperation with other actors."[30] For internet platforms, remedy mechanisms may include disclosing a clear process by which users can challenge terms of service enforcement actions the company has taken against them and appeal for reinstatement of their content or account. Companies should also make sure that their remedy mechanisms are broad enough to cover a range of complaints that users may submit.

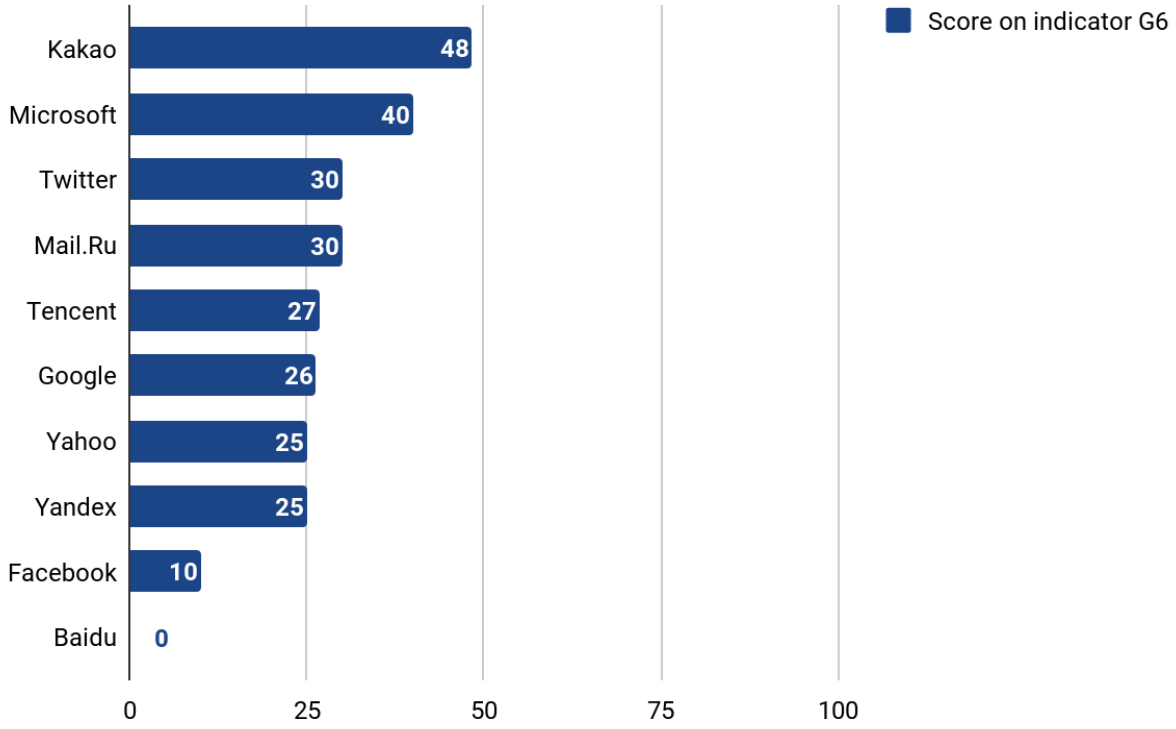
Offline power structures are often replicated online, and in maintaining remedy mechanisms, platforms should therefore seek to ensure that marginalized voices are heard. However, this is not currently the case. For example, several Rohingya activists have reported that their **Facebook** accounts and/or content have been repeatedly censored.[31] The content at issue ranged from news about military atrocities, news about military action in Rakhine state, and even a poem about refugees fleeing military violence. "I have deactivated my account in frustration," one individual told the *Daily Beast*. These activists rely on platforms like Facebook to help them spread awareness and news of a conflict often neglected by mainstream media, but Facebook's repeated deleting of their content and threats to remove their accounts silences voices speaking out about and documenting atrocities. In cases such as this, the lack of clear grievance and remedy mechanism can exacerbate the original freedom of expression concern, escalating the situation from an instance of censorship to exile from the platform altogether (whether self-imposed or as a result of the company deactivating the user's account). Facebook does allow users to appeal some types of decisions but not all (users cannot appeal decisions to removal individual posts, for example),[32] and descriptions of the appeals process reveal that it is not always straightforward or addressed to the user's actual complaint.[33]

At present, grievance and remedy mechanisms companies offer are totally inadequate to match the enormous influence these platforms wield over freedom of expression.

Figure 6 on the next page illustrates the performance of the ten internet platforms evaluated on indicator G6, which examines company disclosures about their grievance and remedy mechanisms.

[30] UN Guiding Principles, p. 24: http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf
[31] https://www.thedailybeast.com/exclusive-rohingya-activists-say-facebook-silences-them
[32] https://onlinecensorship.org/resources/how-to-appeal
[33] https://www.eff.org/deeplinks/2015/03/facebook-has-clarified-its-policies-how-about-fixing-them

Figure 6: Remedy



As pressure from governments around the world to police content continues to rise, having robust grievance and remedy mechanisms is essential to mitigating harms to freedom of expression and correcting errors. Social and search platforms at present are not putting sufficient effort into grievance and remedy, as the 2017 Index data shows. Notably:

- Of the ten platform companies evaluated, **Kakao** had the highest score, with 58%. While this disclosure was largely due to requirements under South Korean law, Kakao went beyond the legal requirement for compliance by also providing users with an appeals mechanism when content is removed in response to defamation claims. This therefore is not only an example of regulation playing a positive role, but also of a company going above and beyond the minimum legal requirement.

- Only two platform companies—**Kakao** and **Mail.Ru**—fully disclosed across all services that their grievance and remedy mechanisms include complaints related to freedom of expression. **YouTube** also received full credit on this element, though **Google Search** received partial credit (G6, Element 2, see appendix).

- **Twitter** was the only platform company to receive any credit for disclosing the number of complaints it receives related to freedom of expression. However, it received partial credit, as the copyright notices section of its Transparency Report only includes numbers of DMCA takedowns and not other forms of complaints involving freedom of expression.[34]

Having adequate remedy mechanisms in place to address freedom of expression harms is especially important in light of increasing government pressure for platforms to regulate extremist content. This has come in the form of new laws, such as Germany's NetzDG law, mentioned earlier in this submission, as well as from governments publicly calling on platforms to step up efforts to remove extremist content. Company attempts to ensure compliance appears to be causing them to err on the side of censorship—the consequences of which can be severe as demonstrated by previously cited examples such as the silencing of Rohingya voices in Myanmar or Syrian opposition activists. Remedy mechanisms can mitigate the harm—for example, so that thousands of videos depicting violence in Syria can be restored and preserved for potential use in future war crime prosecutions.[35] Effective grievance reporting mechanisms may also help to inform governments and companies about the consequences of certain types of policing mechanisms, in addition to helping companies and policymakers develop alternative policies and practices that result in less collateral censorship.

Conclusion

Although the information that internet platforms currently disclose sheds some light on the state of content regulation in the digital age, overall, users and advocates are still left in the dark. The lack of company transparency is especially concerning in a political context where governments and other stakeholders concerned about hate speech, extremism, and other malicious behavior online are pushing for greater content restrictions. The past year has seen a wide variety of efforts to push companies to more vigorously police their platforms, from proposals for automated systems to detect and filter copyright,[36] and remove extremist content[37] to combating hate and abuse.[38] These proposals have different scopes and aims, but until companies are more transparent about their own policies, and how they respond to third party requests, it is unclear what further impact the newly proposed measures may have on users' freedom of expression rights. It is difficult to gauge the potential impact to freedom of

[34] https://transparency.twitter.com/copyright-notices/2015/jul-dec
[35] http://www.bbc.com/news/technology-41023234
[36] https://www.eff.org/deeplinks/2017/10/digital-rights-groups-demand-deletion-unlawful-filtering-mandate-proposed-eu
[37] https://cdt.org/blog/pressuring-platforms-to-censor-content-is-wrong-approach-to-combatting-terrorism/
[38] https://www.theverge.com/2017/10/14/16410348/twitter-boycott-new-rules-enforcement-aggressive-stance-harassment-jack-dorsey

expression when little data about content restrictions is available, when companies themselves do not appear to be conducting comprehensive impact assessments, and when users whose voices are silenced do not have adequate grievance and remedy mechanisms through which to lodge complaints and restore their content.

Meanwhile, freedom of expression advocates warn that the trend towards privatizing censorship - leaving the rule-making and enforcement to companies with minimal government or court involvement - raises significant freedom of expression risks, and that these efforts come with significant negative consequences.[39] RDR agrees with those who argue that legislation that increases intermediary liability will not advance freedom of expression but rather would create additional risks for violating it. On the other hand, there has been relatively little discussion among stakeholders about laws or incentives that would increase transparency about how expression is governed online by companies and governments. Measures requiring or incentivizing internet platforms to provide greater transparency to users and the public, to conduct risk assessments on potential threats their services and policies may have to freedom of expression, and offering robust remedy mechanisms, could potentially improve accountability among all actors and would also help to inform better solutions to the real problem of malicious content and behavior online.

Corporate transparency, however, must be matched by commitment from governments to disclose comprehensive data about the volume and nature of requests being made to companies. Governments should similarly conduct human rights impact assessments to identify potential adverse impacts that may be caused by the enforcement of laws targeting content on internet platforms. Finally, governments should ensure that effective legal and other types of remedies are available to those whose freedom of expression rights are infringed when companies attempt to comply with laws and other requirements to police content.

[39] https://cdt.org/blog/pressuring-platforms-to-censor-content-is-wrong-approach-to-combatting-terrorism/

Recommendations

For companies:

- Improve transparency and accountability about all types of third-party requests to restrict content or user accounts—government requests as well as requests by private individuals and organizations. To the maximum extent possible under the law, companies should publish comprehensive information (including transparency reports) related to the following types of third-party requests:
    - Process for responding to all types of third-party requests to restrict content, access, or service;
    - Data about government requests to restrict content, access, or service;
    - Data about private requests for content restriction.
- Publish data on a regular basis about the volume and nature of content removals and account restrictions that the company makes to enforce its terms of service.
- Conduct regular assessments to determine the impact of the company's products, services, and business operations on users' freedom of expression and privacy. Companies should carry out human rights impact assessments on a regular basis and continue to assess the potential impact of their products and services after the initial rollout. Companies should also pay close attention to the implications of their terms of service enforcement policies when conducting HRIAs.
- Conduct meaningful outreach to different stakeholder groups in order to learn how their services may be negatively impacting users in different communities, and identify ways to mitigate these harms.
- Establish effective grievance and remedy mechanisms, and clearly indicate that these mechanisms can be used to raise concerns related to potential or actual violations of freedom of expression and privacy.

For governments:

- Instead of passing laws that would increase intermediary liability on internet platforms, governments should perhaps consider regulation requiring greater transparency from internet platforms. This would not only help to advance freedom of expression on these platforms, but could also help to address ongoing concerns of their use in potential misinformation and manipulation of public opinion.[40]

- Consider the potential for regulation requiring companies to conduct risk assessments on potential threats their services and policies may have to freedom of expression, and offer robust remedy mechanisms.

- Ensure that laws and regulations allow companies to be transparent and accountable with users about how they receive and handle government requests to restrict content.

- Conduct human rights impact assessments on all proposed and existing laws related to the regulation of online platforms to identify and mitigate potential violations of users' rights to freedom of expression.

- Work with companies to establish effective legal and other types of grievance and remedy mechanisms for people whose right to freedom of expression has been infringed in the course of policing platforms for malicious content.

[40] https://www.nytimes.com/2017/10/29/business/facebook-misinformation-abroad.html

Appendix: Full text of relevant 2017 Corporate Accountability Index indicators

The full methodology, including definitions and research guidance, used for the 2017 Index can be downloaded at: https://rankingdigitalrights.org/wp-content/uploads/2016/09/2017Indexmethodology.pdf

Raw research data for all indicators, along with the full 2017 Index report can be downloaded at: https://rankingdigitalrights.org/index2017/download/

Full text of indicators discussed in this submission is duplicated below, plus links to indicator research guidance and definitions as well as links to the relevant 2017 Index results for each indicator.

Indicators F3-F7 assess transparency about company actions affecting freedom of expression.
Indicator G4 assesses disclosure of impact assessment.
Indicator G6 assesses disclosure of grievance and remedy mechanisms.

F3. Process for terms of service enforcement

https://rankingdigitalrights.org/2017-indicators/#F3

The company should **clearly disclose** the circumstances under which it may restrict **content** or **user accounts**.

Results: https://rankingdigitalrights.org/index2017/indicators/#F3

Elements:

1. Does the company **clearly disclose** what types of content or activities it does not permit?

2. Does the company **clearly disclose** why it may **restrict a user's account**?

3. Does the company **clearly disclose** information about the processes it uses to identify **content** or **accounts** that violate the company's rules?

4. Does the company **clearly disclose** whether any government authorities receive priority consideration when flagging content to be restricted for violating the company's rules?

5. Does the company **clearly disclose** whether any private entities receive priority consideration when flagging content to be restricted for violating the company's rules?

6. Does the company **clearly disclose** its process for enforcing its rules?

7. Does the company provide clear examples to help the user understand what the rules are and how they are enforced?

F4. Data about terms of service enforcement

https://rankingdigitalrights.org/2017-indicators/#F4

The company should **clearly disclose** and regularly publish data about the volume and nature of actions taken to restrict content or accounts that violate the company's rules.

Results: https://rankingdigitalrights.org/index2017/indicators/#F4

Elements:

1. Does the company **clearly disclose** data about the volume and nature of content and accounts restricted for violating the company's rules?

2. Does the company publish this data at least once a year?

   Can the data published by the company be exported as a **structured data** file?

F5. Process for responding to third-party requests for content or account restriction

https://rankingdigitalrights.org/2017-indicators/#F5

The company should **clearly disclose** its process for responding to **government requests** (including judicial orders) and **private requests** to remove, filter, or restrict **content** or **accounts**.

Results: https://rankingdigitalrights.org/index2017/indicators/#F5

Elements:

1. Does the company **clearly disclose** its process for responding to **non-judicial government requests**?

2. Does the company **clearly disclose** its process for responding to **court orders**?

3. Does the company **clearly disclose** its process for responding to **government requests** from foreign jurisdictions?

4. Does the company **clearly disclose** its process for responding to **private requests**?

5. Do the company's explanations **clearly disclose** the legal basis under which it may comply with **government requests**?

6. Do the company's explanations **clearly disclose** the basis under which it may comply with **private requests**?

7. Does the company **clearly disclose** that it carries out due diligence on **government request**s before deciding how to respond?

8. Does the company **clearly disclose** that it carries out due diligence on **private requests** before deciding how to respond?

9. Does the company commit to push back on inappropriate or overbroad **requests made by governments**?

10. Does the company commit to push back on inappropriate or overbroad **private requests**?

11. Does the company provide clear guidance or examples of implementation of its process of responding to **government requests**?

12. Does the company provide clear guidance or examples of implementation of its process of responding to **private requests**?

F6. Data about government requests for content or account restriction

https://rankingdigitalrights.org/2017-indicators/#F6

The company should regularly publish data about **government requests** (including judicial orders) to remove, filter, or restrict **content** or **accounts**.

Results: https://rankingdigitalrights.org/index2017/indicators/#F6

Elements:

1. Does the company break out the number of requests it receives by country?

2. Does the company list the number of accounts affected?

3. Does the company list the number of pieces of content or URLs affected?

4. Does the company list the types of subject matter associated with the requests it receives?

5. Does the company list the number of requests that come from different legal authorities?

6. Does the company list the number of requests with which it complied?

7. Does the company publish the original requests or disclose that it provides copies to a **public third-party archive**?

8. Does the company reports this data at least once a year?

9. Can the data be exported as a **structured data** file?

F7. Data about private requests for content or account restriction

https://rankingdigitalrights.org/2017-indicators/#F7

The company should regularly publish data about **private requests** to remove, filter, or restrict access to **content** or **accounts**.

Results: https://rankingdigitalrights.org/index2017/indicators/#F7

Elements:

1. Does the company break out the number of requests it receives by country?
2. Does the company list the number of **accounts** affected?
3. Does the company list the number of pieces of content or URLs affected?
4. Does the company list the reasons for removal associated with the requests it receives?
5. Does the company describe the types of parties from which it receives requests?
6. Does the company list the number of requests it complied with?
7. Does the company publish the original requests or disclose that it provides copies to a **public third-party archive**?
8. Does the company report this data at least once a year?
9. Can the data be exported as a **structured data** file?
10. Does the company **clearly disclose** that its reporting covers all types of **private requests** that it receives?

G4. Impact assessment

https://rankingdigitalrights.org/2017-indicators/#G4

The company should conduct regular, comprehensive, and credible due diligence, such as **human rights impact assessments**, to identify how all aspects of its business affect freedom of expression and privacy and to mitigate any risks posed by those impacts.

Results: https://rankingdigitalrights.org/index2017/indicators/#G4

Elements:

1. As part of its decision-making, does the company consider how laws affect freedom of expression and privacy in jurisdictions where it operates?

2. Does the company regularly assess freedom of expression and privacy risks associated with existing products and services?

3. Does the company assess freedom of expression and privacy risks associated with a new activity, including the launch and/or acquisition of new products, services, or companies or entry into new markets?

4. Does the company assess freedom of expression and privacy risks associated with the processes and mechanisms used to enforce its terms of service?

5. Does the company conduct additional evaluation wherever the company's risk assessments identify concerns?

6. Do **senior executives** and/or members of the company's board of directors review and consider the results of assessments and due diligence in their decision-making?

7. Does the company conduct assessments on a regular schedule?

8. Are the company's assessments assured by an external third party?

9. Is the external third party that assures the assessment accredited to a relevant and reputable human rights standard by a credible organization?

G6. Remedy

https://rankingdigitalrights.org/2017-indicators/#G6

The company should have **grievance** and **remedy** mechanisms to address users' freedom of expression and privacy concerns.

Results: https://rankingdigitalrights.org/index2017/indicators/#G6

Elements:

1. Does the company **clearly disclose** its processes for receiving complaints?

2. Does the company **clearly disclose** that its process includes complaints related to freedom of expression and privacy?

3. Does the company **clearly disclose** its process for responding to complaints?

4. Does the company report on the number of complaints received related to freedom of expression and privacy?

5. Does the company provide clear evidence that it is responding to complaints?