Making Transparency Meaningful

A Framework for Policymakers

December 2021

CENTER FOR DEMOCRACY & TECHNOLOGY

The **Center for Democracy & Technology** (CDT) is a 25-year-old 501(c)3 nonpartisan nonprofit organization working to promote democratic values by shaping technology policy and architecture. The organisation is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

Making Transparency Meaningful

A Framework for Policymakers

Authors

Caitlin Vogus

Emma Llansó

December 2021

Table of Contents

Introduction

Transparency is everywhere in policy debates over the responsibilities of technology companies and how best to regulate them. And for good reason. Tech companies have promised greater transparency, and lawmakers in the United States,[1] Europe,[2] and elsewhere have proposed legislation that would enhance or require transparency, often at the urging of civil society. Transparency can enhance public understanding of how technology companies operate and make them more accountable, whether through public pressure or legal constraints. It is offered as part of a solution to difficult problems raised by technology, from combating the spread of mis- and disinformation to reining in government surveillance to addressing discriminatory online advertising and more.

But what exactly do we mean when we talk about transparency when it comes to technology companies like social networks, messaging services, and telecommunications firms? Transparency can take a variety of forms, and different stakeholders will find different types of transparency useful or important. And different forms of transparency give rise to varying technical, legal, and practical challenges.

This paper sets forth a conceptual framework for transparency about practices that affect users' speech, access to information, and privacy from

1 Members of Congress have introduced numerous transparency bills, such as the Algorithmic Justice and Online Platform Transparency Act, the Platform Accountability and Consumer Transparency Act, and the Social Media Disclosure And Transparency of Advertisements Act of 2021.

2 In Europe, lawmakers have introduced extensive, multifaceted transparency obligations within the draft regulation known as the Digital Services Act and additional voluntary frameworks.

government surveillance.[3] It maps and briefly describes current and past efforts at transparency in four distinct categories:

1. Transparency reports that provide aggregated data and qualitative information about moderation actions, disclosures, and other practices concerning user generated content and government surveillance;

2. User notifications about government demands for their data and moderation of their content;

3. Access to data held by intermediaries for independent researchers, public policy advocates, and journalists; and

4. Public-facing analysis, assessments, and audits of technology company practices with respect to user speech and privacy from government surveillance.

The purpose of the framework is to delineate the different ways that policymakers, civil society, the private sector, and the public are discussing technology company transparency in order to provide greater clarity about the potential benefits and tradeoffs that come with each form of transparency. Discussions of each of these four categories of transparency mechanisms often happen in parallel, as if each is entirely separate from the others. However, efforts to improve transparency must appreciate how the different forms are linked and where they differ, as well as the challenges and tradeoffs in enhancing each form of transparency through voluntary and regulatory interventions.

For example, various forms of transparency could be helpful in combating the spread of disinformation online. Policymakers and the public could gain a better general understanding of how disinformation spreads, and who it affects, through company transparency reporting, but answering specific empirical questions about the patterns and consequences of disinformation will necessitate providing independent researchers with access to data to conduct their research. Both transparency reporting and researcher access to data approaches will need to grapple with defining "disinformation," but transparency reporting requirements may also need to resolve how companies should count their content

3 Numerous other topics – such as safety, supply chains and labor practices, sustainability and the environment, and employee diversity – have also been subject to calls for increased transparency by technology companies. These topics, while important, are beyond the scope of this framework.

moderation actions on disinformation, while mandated researcher access to data about online disinformation may need to resolve how to provide access to sensitive data while preserving user privacy. More targeted interventions, to help individual users navigate and debunk disinformation online, will require user-centric transparency and careful thinking about what kind of information is useful and actionable to users. And any third-party analysis or evaluation of company performance in combating disinformation would need to proceed from clear and objective criteria – likely informed by the findings from transparency reports, user notifications, and independent research.

The framework in this paper provides a structure for understanding the big picture of technology company transparency, and how to approach the critical decisions that must be made if we are to achieve meaningful transparency by, and about, technology companies.

Transparency Reports

Many tech companies produce regular *transparency reports* about how their actions affect the speech, access to information, and privacy of their users with respect to government surveillance. These are public reports that may include aggregate data and qualitative information about the reporting entity's operations. One goal of transparency reporting is to increase companies' accountability by enhancing public understanding about how their services impact users, communities, and other stakeholders. While transparency reports published by tech companies are most common, other entities, including governments, may also issue transparency reports on issues that impact users' speech and privacy. Transparency reports can provide a better understanding of the overall environment for online speech and participation on a particular service and allow users and others to see how different content moderation practices and government demands for content restriction have changed over time.

////

Current State of Transparency Reporting

As of July 2021, 88 technology companies – including telecoms, social networks, search engines, and e-commerce companies – had issued transparency reports concerning free expression and user privacy.[4] In 2010, Google published the first transparency report, about government demands for user data. In the years that followed, other tech companies began publishing similar transparency reports, and information about government demands for user data is now the most common element of tech company transparency reports.

4 *Transparency Reporting Index*, Access Now (last updated July 2021).

Transparency reporting on other topics grew from these initial reports on government demands for user data. Now, transparency reports may also cover: government demands for content removals;[5] content removals and other content moderation by companies under their own terms of service (also known as "content policy enforcement");[6] advertising (including ad libraries);[7] copyright and trademark enforcement and other legal requests;[8] and network shutdowns and disruptions.[9]

Content Policy Enforcement Transparency Reports

Voluntary reporting about content removals and other content moderation under companies' terms of service became more common starting in 2018 following publication of the Santa Clara Principles on Transparency and Accountability in Content Moderation, which promote due process and transparency as ways of ensuring that companies' enforcement of their content guidelines is more "fair, unbiased, proportional, and respectful of users' rights."[10] The data in these reports varies because companies' content policies – and enforcement of those policies – differ widely. Generally, however, existing transparency reports about content policy enforcement provide data about the volume and nature of content removed. They sometimes also provide information about how violating content was detected or reported and the number and outcomes of appeals.

Recently, some countries have required (or strongly encouraged) online service providers to publish transparency reports about the removal of illegal content. For example, the German Network Enforcement Act (NetzDG) requires certain internet companies to

5 E.g., Twitter's transparency report on Removal Requests reports on "legal demands to remove content from Twitter and Periscope, and other requests to remove content based on local law(s) from around the world."

6 E.g., TikTok publishes a quarterly Community Guidelines Enforcement Report.

7 E.g., Google publishes a transparency report on Political Advertising in the United States, which includes an ad library.

8 E.g., GitHub's annual transparency report includes a section on takedowns under US copyright law; similarly, Microsoft's Content Removal Requests Report includes sections on both copyright removal requests and "right to be forgotten" requests.

9 E.g., telecommunications companies Telefónica and Telenor disclose the number of network shutdown demands they receive, and AT&T provides "partial disclosure" of such demands. *2020 Ranking Digital Rights Corporate Accountability Index* at F10. Network shutdown (telecommunications companies) (last visited Dec. 7, 2021).

10 *The Santa Clara Principles on Transparency and Accountability in Content Moderation* (last visited Nov. 21, 2021) [hereinafter "Santa Clara Principles"]. Content policy enforcement reports are published by Facebook, Twitter, YouTube, Reddit, Pinterest, SnapChat, TikTok, Twitch, and many other online intermediaries.

publish semi-annual transparency reports with specific information about their content moderation practices with respect to content that is illegal under German law. Several companies have published separate transparency reports pursuant to NetzDG, in addition to their voluntary transparency reports.[11]

Government Transparency Reports

A few governments also publish selective transparency reports relevant to the speech and surveillance of users of telecommunications or online services. For example, in the United States, the Administrative Office of the United States Courts and the Director of National Intelligence are required by law to publish annual reports on the use of certain surveillance authorities.[12] In another example, in Europe, the European Commission publishes reports monitoring the implementation of the Code of Conduct for Countering Illegal Hate Speech Online, which includes aggregate data about reports made to technology companies under the Code.[13] In general, however, most government entities around the world have no habit or legal requirement to produce publicly available reports about governmental efforts to restrict online speech or obtain the information or data of users of telecommunications or online services.

////

Enhancing Transparency Reporting: Considering Tradeoffs

Would increasing voluntary transparency reporting, or mandating transparency reporting, enhance technology companies' accountability to the public?

Some experts and advocates are skeptical of the benefits of transparency reporting, arguing that self-reported aggregate data often do not offer true insight into how content practices work and therefore cannot be used to hold companies accountable for their

11 Heidi Tworek & Paddy Leerssen, *An Analysis of Germany's NetzDG Law* at 10 (Apr. 15, 2019) (Appendix with links to NetzDG transparency reports from Google, Facebook, and Twitter).

12 *See* 18 U.S.C. § 2519; 18 U.S.C. § 1873(a)(2) & (b). The information in these reports is limited, however, and some of it has been criticized as potentially inaccurate. *See* Albert Gidari, *Wiretap Numbers Don't Add Up*, Just Security (July 6, 2015); Albert Gidari, *The Government's Wiretap Orders Still Don't Add Up*, Just Security (July 17, 2015); Albert Gidari, *Wiretap Reports Not So Transparent*, Ctr. for Internet and Society (Jan. 26, 2017).

13 Barbora Bukovská, *The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An analysis of freedom of expression implications*, Article 19 at 7-10 (May 7, 2019).

decisions and impacts.[14] Transparency reports do not typically reveal the underlying content discussed in them, which some argue is necessary for accountability-enhancing transparency.[15] Moreover, because companies control collection and reporting of the data in many transparency reports, others have questioned whether that data is accurate and how it can be validated.[16] Even when reports are issued by governments, the data included may not be particularly meaningful; for example, the European Commission's reports on the Code of Conduct for Countering Illegal Hate Speech Online have been criticized for focusing on the rate and speed of content removals rather than an analysis of the type of content removed.[17] These concerns may be partially addressed through efforts to incentivize or mandate transparency reports that contain specific data which would provide meaningful transparency. However, the value of transparency reporting has limits, and they are one method among several of improving tech company accountability.

What qualitative and quantitative data and information should be disclosed in transparency reports?

Publishers of different types of transparency reports – or lawmakers who would mandate them – must determine the specific data that should be included. It is not feasible for tech companies to track and report all possible data, and such an approach would raise concerns about user privacy and the costs imposed on smaller and newer companies.

In making this determination, the first consideration is what data can be collected and reported. If a company's or government's systems and processes are not designed to track particular data, it will not be able to report it. While it may sometimes be possible to redesign these systems and processes to allow for tracking and reporting of specific data, in other cases it will not. For example,

---

14 *See, e.g.*, Ethan Zuckerman, *I read Facebook's Widely Viewed Content Report. It's really strange.*, ...My heart's In Accra Ethan Zuckerman's online home, since 2003 (Aug. 18, 2021); Laura Edelson, *Facebook's political ad spending numbers don't add up*, Medium (Oct. 12, 2020); Davey Alba, Catie Edmondson & Mike Isaac, *Facebook Expands Definition of Terrorist Organizations to Limit Extremism*, N.Y. Times (Sept. 17, 2019) (quoting evelyn douek).

15 *See* Zuckerman, *supra* n.14.

16 *See* Eric Goldman, *RightsCon 2021 Lightning Talk: Validating & Enforcing Transparency Reports*, YouTube (June 7, 2021).

17 Bukovská, *supra* n.13 at 9. *See also* Jens-Henrik Jeppesen, First report on the EU Hate Speech Code of Conduct shows need for transparency, judicial oversight, and appeals, Ctr. for Democracy & Tech. (Dec. 12, 2016).

companies that offer end-to-end-encrypted services will be unable to capture certain data concerning user-generated content that is technically inaccessible to them. In other cases, the way a service operates will make it profoundly difficult to collect certain data, even if it is not technically impossible. For example, Wikipedia may not be able to capture and report aggregate data about content removed by its hundreds of thousands of volunteer editors under its content policies. Finally, decisionmakers should understand that, in some cases, companies are legally prohibited from disclosing particular data or information.[18]

The intended audience of a transparency report is another important consideration in deciding what data should be provided and in what format. An expert audience may appreciate transparency reports with granular and deeply technical data, whereas lay audiences will find reports with narrative information and additional explanations of quantitative data more accessible.

Civil society organizations have provided guidance on the data that voluntary transparency reports should include. While these recommendations may be a useful starting point for policymakers considering requiring tech companies to publish transparency reports, they are not model legislation and should not be incorporated wholesale into proposals that would mandate transparency reporting. Examples of civil society guidance on voluntary transparency reports include:

- New America's Open Technology Institute's two transparency reporting toolkits, one focused on government requests for user data[19] and the other on content takedown reporting,[20] which recommend best practices for transparency reports and the type and granularity of data that internet and telecommunications companies should provide;

18 For example, prior to the passage of the USA FREEDOM Act in 2015, companies' ability to report on national security letters and orders they received under the Foreign Intelligence Surveillance Act (FISA) was severely restricted. The USA FREEDOM Act loosened these restrictions, though reporting about NSLs and FISA orders is still limited. Companies' ability to disclose information about network shutdowns and other disruptions may also be limited by law.

19 Liz Woolery, Ryan Budish, & Kevin Bankston, *The Transparency Reporting Toolkit: Reporting Guide & Template for Reporting on U.S. Government Requests for User Information*, New America & The Berkman Klein Center For Internet & Society (Dec. 2016).

20 Spandana Singh & Kevin Bankston, *The Transparency Reporting Toolkit: Content Takedown Reporting*, New America (Oct. 24, 2018).

- The Center for Democracy & Technology and Global Network Initiative's recommendations about the data and information that governments should report concerning surveillance and content removal and restriction, as well as the information that governments should permit tech companies to disclose;[21] and

- The Santa Clara Principles, which set forth basic standards for transparency reporting on content policy enforcement and specify the minimum data that these transparency reports should include.[22]

How should data in transparency reports be categorized and counted?

Publishers of transparency reports must also decide how to categorize and count the data they report. Even reports on the same topic may categorize and count data differently. For example, in reports about government demands for user data, some companies provide the numbers of demands for each separate category of legal process used – such as pen registers, wiretaps, or search warrants – while others report numbers for combined categories of legal processes, and others still report only a single number for all government demands for user data, regardless of the type of legal process used. Similar issues arise in transparency reports about content policy enforcement. Many companies organize these reports around the categories in their content policies, which differ from company to company.

In addition, even when companies use similar categories, they may not count data in the same way.[23] For example, if a post is removed for violating multiple provisions of a company's content policy, it could be counted as a single removal or separately under each provision for which it was removed.

---

21 Emma Llansó & Susan Morgan, *Getting Specific About Transparency, Privacy, and Free Expression Online*, Ctr. for Democracy & Tech. (Nov. 5, 2014).

22 Santa Clara Principles, *supra* n.10.

23 *See* Daphne Keller, *Some Humility About Transparency*, Ctr. for Internet & Society (Mar. 19, 2021) (linking to a Google Doc in which Keller sets forth a partial list of logistical and operational questions that arise when building a transparency report); see also Andrew Puddephatt, *Letting the sun shine in: transparency and accountability in the digital age*, UNESCO at 15 (2021) (raising questions about transparency reporting, including "how is an 'item' of data defined? How is a URL containing thousands of illegal images counted?").

Policymakers could specify how data regarding lawful orders for content restriction or user data are to be categorized and counted in transparency reports, but doing so may influence how companies conduct the processes on which they are reporting and may limit the development of new methods of categorization and counting that could provide clearer or more meaningful information.

Is standardization of transparency reports possible or desirable?

Some critics of transparency reports have noted that a lack of standardization across reports makes comparison between tech companies difficult. Calls for standardization of transparency reports must consider two questions: Is standardization possible? And is it desirable? As discussed above, companies report different data in their transparency reports, and, even when they appear to report the same data, they may categorize and count it differently. Standardizing precisely how to categorize and count data can be difficult, especially given differences in the services offered, content moderation rules and processes, advertising models, and other facets of companies. Moreover, most categories of speech that a company might restrict lack a standardized definition that applies across cultural and national contexts; there is not, for example, a standardized definition of "extremist content" or "sexual imagery" that could be applied across all services, even if many companies restrict those general categories of content.

Transparency reporting standardization may also have unintended negative side effects. Requiring tech companies to provide specific metrics in transparency reports could stifle innovation in reporting and sharing of other data that turn out to be more meaningful for a particular service.[24] Highly prescriptive transparency requirements may also force or encourage intermediaries to standardize their content policies or content moderation practices, creating a more homogeneous online environment and decreasing the variety of options for online services from which users can choose.

Does mandatory transparency reporting about content policy enforcement incentivize over-removal of speech or otherwise influence content enforcement?

Requiring companies that host user-generated content to report certain data about content removals under their content policies

24 Spandana Singh, *A Spotlight on Transparency: An Overview of How the Practice of Transparency Reporting Has Emerged Across Different Industries*, New America at 12-17 (Apr. 2020).

may encourage them to remove speech, even if it is legal and does not violate the companies' content policies, in at least two ways. First, services that must comply with reporting obligations concerning content moderation may respond by adopting "simpler, blunter content rules" that are either overly broad or narrow to make it easier to classify and explain their decisions.[25] Second, a service may feel pressured to report high numbers of removals of certain kinds of speech and respond by removing more speech than is actually prohibited under its content policy. For example, a company that publishes the number of posts removed as terrorist content may err on the side of removing gray-area content that does not actually advocate for terroristic violence so its transparency report will show a high number of removals under that category. As a result, content that is in the public interest – such as news reports about terrorism – may be overremoved.

Mandatory content policy enforcement transparency reporting may also encourage companies to devote a disproportionate amount of resources to the types of content contained in the report and fewer resources to other types of problematic content on which they are not required to report. For example, if a company must report the number of content items it removes as a result of its hate speech policy but not as a result of a policy against disinformation, it may devote more resources to hate speech detection and removal and fewer to disinformation. Similarly, a requirement that companies report data such as the length of time it took them to remove particular content may incentivize them to make content removal decisions faster, even if that means more errors.

What is the impact of transparency reporting on smaller and startup companies?

Transparency reporting can be expensive and labor intensive. Detailed and far-reaching requirements for transparency reporting, in particular, can negatively impact smaller tech companies, entrench dominant companies, and decrease competition and pluralism among providers. As a result, it may be necessary to exempt smaller or newer companies from transparency reporting mandates or make distinctions about what data or how much data they must report. These distinctions can be based on metrics such as the age of the company, number of employees, revenues, or consumer usage, with benefits and downsides to each metric that

25 Keller, *supra* n.23; *see also* Puddephatt, *supra* n.23 at 15 (raising the question of whether "adoption of rules for disclosing content moderation make companies adopt simpler rules that do not take account of nuance").

can be used.[26] Another approach would be to implement high-level principles for transparency, rather than detailed metrics, that all companies could meet.[27]

Can transparency reporting be mandated in the United States consistent with the First Amendment?

American lawmakers considering mandates for transparency reporting should examine whether doing so is consistent with the First Amendment. In general, strict scrutiny applies to statutes that compel speech by private speakers, including "not only to expressions of value, opinion, or endorsement, but equally to statements of fact the speaker would rather avoid."[28] At least one court has struck down a state law that would have required online platforms to publish certain information about political advertising,[29] a holding which could be extended to transparency reporting. In addition, lawmakers should consider whether requiring content policy enforcement reports, which would require hosts of user generated content to disclose data about their decisions to publish or remove content, would impinge on their First Amendment right to exercise editorial discretion over the content they host.[30]

26 Eric Goldman & Jess Miers, *Regulating Internet Services by Size*, CPI Antitrust Chronicle, Santa Clara Univ. Legal Studies Research Paper (May 2021).

27 *See* Puddephatt, *supra* n.23 at 2.

28 *Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Boston*, 515 U.S. 557, 573 (1995).

29 *Washington Post v. McManus*, 944 F.3d 506 (4th Cir. 2019).

30 *Herbert v. Lando*, 441 U.S. 153 (1979); *Miami Herald v. Tornillo*, 418 U.S. 241 (1974).

User Notifications

Technology companies may notify users about a variety of activities that affect their speech, access to information, and privacy. Three types of user notifications that most strongly impact – and can help protect – user privacy and speech are:

1. *Government demands for user data;*
2. *Legal demands for content removals or restrictions; and*
3. *Content moderation decisions by companies.*

Notice about government demands for user data gives the user the opportunity to challenge the release of their data to the government and helps shed light on the often opaque processes of government surveillance. Similarly, notice about legal demands for content removals or restrictions gives users the opportunity to challenge those demands and reveals how governments and civil litigants obtain content takedowns or other restrictions. Notice about content moderation can educate users about intermediaries' content policies and reveal how and why intermediaries moderate content. All of these forms of user notifications inform public opinion and policymaking, helping hold governments accountable for their online surveillance and censorship activity and intermediaries for their content moderation practices.

////

Current Approaches to User Notifications

Government demands for user data

Governments around the world may demand data about users from technology companies, including users' content and non-content data such as traffic data as well as subscriber and billing information. Many tech companies have a policy of informing users of government demands for their data before turning it over unless they are prohibited from doing so by law or by other limited exceptions to their policies, such as emergency circumstances that threaten serious injury or death.[31]

In the United States, certain laws or judicial orders can prohibit a company from notifying users about a government demand for their data or require that they delay providing such notice. For example, the federal wiretap statute, Title III, prohibits a provider of wire or electronic communication service from disclosing the existence of a wiretap (an ongoing form of surveillance).[32] The Stored Communications Act (SCA) permits the government to obtain some forms of electronic communications data without itself providing notice to the targeted user if the government obtains a warrant, or with delayed notice if it obtains a subpoena or court order under 18 U.S.C. § 2703(d) and meets certain statutory criteria. The SCA further authorizes issuance of a gag order precluding the company that receives the warrant, subpoena, or order from providing notice to the targeted user in certain circumstances.[33] The SCA also authorizes the FBI to issue a gag with a National Security Letter (NSL), a type of administrative subpoena, precluding the recipient from disclosing the existence of the NSL, if the FBI certifies that certain statutory criteria are met.[34] Providers are not permitted to

31 *See* Nate Cardozo et al., *Who Has Your Back? Government Data Requests 2017*, Elec. Frontier Found. (July 10, 2017) (evaluating twenty-six major technology companies on their policy and advocacy positions concerning "handing data to the government," including user notifications).

32 18 U.S.C. § 2511(2)(a)(ii).

33 18 U.S.C. § 2703(b)(1); *id.* § 2705. The SCA also permits the government to obtain non-content records without notice and to obtain a gag order precluding the provider of electronic communication service or remote computing service from notifying the affected user. *See* 18 U.S.C. § 2703(c); *id.* § 2705(b). Department of Justice guidelines limit the circumstances under which it will seek a gag order pursuant to § 2705(b) and limit gag orders' duration to one year other than in exceptional circumstances. *See* *Memorandum from Rod J. Rosenstein, Deputy Attorney General, to Heads of Dep't Law Enforcement Components, Dep't Litigating Components, Director, Exec. Office for U.S. Attorneys, All United States Attorneys* (Oct. 19, 2017). However, the Department's policy is intended "only to improve the internal management of the Department of Justice," and the Department expressly contemplates that orders of a longer duration may be necessary. *Id.* at 1 n.1 & 2 n.3.

34 18 U.S.C. § 2709(c); *see also* 18 U.S.C. § 3511.

disclose the fact that they have received orders to produce data pursuant to the Foreign Intelligence Surveillance Act (FISA), and Section 604(a) of FISA[35] permits providers to report only statistical information on the number of demands they receive under particular authorities. While these legal provisions can prevent or delay a company from notifying users of government demands for their data, some tech companies have a policy of providing notice after a legal prohibition on notice is lifted or expires.[36]

Legal demands for content removals or restriction

Governments may also demand that companies that host user generated content remove or otherwise restrict content (such as by geoblocking it) because it is allegedly illegal. In addition, private parties may demand that hosts remove or restrict content based on claims that it violates civil law, such as for defamation. Both governments' and private parties' legal demands for content removals or restrictions are often made by serving a court order or other legal authority on the host. A few hosts have a policy of informing users of legal demands for removal or restriction of their content unless they are prohibited from doing so by law, certain narrow emergency circumstances apply, or notice would be futile or ineffective.[37]

Some governments may also seek the removal or restriction of content that is not illegal but allegedly violates a host's content policies. In such cases, through Internet Referral Units or other government entities, the government notifies a host that particular content violates the host's content policy, and the host may remove or restrict it pursuant to its content policy.[38] These governmental efforts to leverage hosts' content policies to obtain removal of speech that is not illegal have been criticized both for allowing extra-legal government censorship and for their lack of transparency, since hosts and governments rarely notify users when their content has been removed pursuant to the host's content

35 50 U.S.C. § 1874.

36 *See* Cardozo et al., *supra* n.31.

37 *See* Andrew Crocker et al., *Who Has Your Back? Censorship Edition 2019*, Elec. Frontier Found. (June 12, 2019) (evaluating sixteen major technology companies on their content moderation policies, including user notifications regarding content takedowns and account suspensions in response to legal demands).

38 Jason Pielemeier & Chris Sheehy, *Understanding The Human Rights Risks Associated With Internet Referral Units*, VOX-Pol (Mar. 26, 2020).

policy as a result of a governmental notification.[39] If users do not receive notifications about these government referrals, they may be unable to challenge their legality and may not even know that they are under government scrutiny.

Content moderation

User notifications concerning content moderation decisions can be divided into three categories or phases of notice: (1) Terms of service and content policies; (2) Notifications of enforcement actions; and (3) Appeals.

Intermediaries that host user-generated content usually notify users about what content is and is not allowed on their services. Intermediaries' terms of service may state what content is allowed or forbidden at a high level of generality,[40] and they often have additional, more detailed content policies, which are sometimes called "community standards."[41] The earliest content policies were relatively simplistic and lacking in detail. However, some – though not all[42] – now consist of a complicated and lengthy system of rules, with exceptions and caveats.[43] Content policies educate users about what they can say and how they should behave on a service, and while some users will intentionally break the rules, others will make a genuine attempt to understand and stay within them. Content policies are generally public and available to anyone, even if they do not have an account on the service.

---

39 *See*, e.g., Tomer Shadmy & Yuval Shany, *Protection Gaps in Public Law Governing Cyberspace: Israel's High Court's Decision on Government-Initiated Takedown Requests*, Lawfare (Apr. 23, 2021) (describing the "invisible handshake" between the Israeli IRU and hosts, through which "[a]ffected individuals are aware that content they posted was removed by an online platform because of incompatibility with the applicable community standards or terms of use; they are not aware of the fact that the platform acted in response to a government takedown request").

40 *See, e.g.*, *Terms of Service*, Facebook at Section 3 (last visited Nov. 29, 2021); *Twitter Terms of Service*, Twitter at Section 3, Twitter (last visited Nov. 29, 2021).

41 *See, e.g.*, *Facebook Community Standards*, Facebook (last visited Nov. 29, 2021); *The Twitter Rules*, Twitter (last visited Nov. 29, 2021).

42 Some content policies provide minimal information. For example, social cataloguing website Goodreads' Community Guidelines consist of eight bullet points with some introductory text and two disclaimers. *Community Guidelines*, Goodreads (last visited Nov. 29, 2021). Its Community Guidelines do not define terms used in it to describe prohibited content, such as "hate speech," "nudity" or "graphic violence."

43 For example, Facebook's content policy prohibiting nudity specifies that it allows images of female breasts if they are "depicting acts of protest, women actively engaged in breast-feeding and photos of post-mastectomy scarring." *Adult Nudity and Sexual Activity*, Facebook (last visited Nov. 29, 2021).

An intermediary may also provide a user with notice when it takes an enforcement action against the user's content or account. Notice may be detailed – including information identifying the content removed, the specific part of the content policy that was violated, how the content was detected and removed, and an explanation of how the user can appeal the decision[44] – or it may be perfunctory. Some intermediaries warn users before taking certain enforcement actions,[45] while others provide notice only after the fact. In addition, whether an intermediary provides a user with notice may depend on the type of enforcement action taken.[46] For example, an intermediary that enforces its content policy using purposefully opaque content moderation practices, such as keeping an account active but allowing only the account holder to view the content they post,[47] may intentionally not notify a user of the enforcement action it takes.

Finally, some intermediaries give users the ability to appeal enforcement decisions, providing a further opportunity to communicate with users about content moderation practices and decisions. The appeals process may allow a user to present new information to the intermediary and ideally results in the intermediary notifying the user of the results of its review with information that is sufficient to allow the user to understand the decision.[48]

////

44 *See Santa Clara Principles*, *supra* n.10. A 2019 report by the Open Technology Institute found that YouTube, Facebook, and Twitter met some though not all of the "notice" recommendations in the Santa Clara Principles. Spandana Singh, *Assessing YouTube, Facebook and Twitter's Content Takedown Policies*, New America (May 7, 2019).

45 For example, Instagram sends a warning to an account at risk of deletion for repeated violations of its Community Standards Enforcement that includes a timeline documenting the account's previous violations. *Account Disable Policy Changes on Instagram*, Instagram (July 18, 2019).

46 Content moderation is not just a binary decision to either take down content or accounts or allow them to remain on a service; depending on how they have designed their service, intermediaries can take a wide variety of actions against violative content, some of which may not be immediately obvious to the user who posted the content. For example, intermediaries may decrease the availability of a post by removing or downgrading its visibility in search results. They may stop recommending certain content or display it less prominently in users' feeds. They may also restrict forwarding or sharing of content. *See* Eric Goldman, *Content Moderation Remedies*, Mich. Tech. L. Rev. (Forthcoming 2021).

47 These opaque content moderation practices are often referred to as "shadowbanning." Gabriel Nicholas, *Spotlight on Shadowbanning*, Ctr. for Democracy & Tech. (Oct. 4, 2021).

48 *See Santa Clara Principles, supra* n.10; A 2019 report by the Open Technology Institute found that YouTube, Facebook, and Twitter met many of the "appeals" recommendations in the Santa Clara Principles. *See* Singh, *supra* n.44.

Improving User Notifications: Considering Tradeoffs

What are the costs and benefits of giving technology companies greater legal authority to disclose government demands for user data?

As explained above, in some cases, tech companies are precluded by law from notifying users about government demands for their data, or they must delay in providing such notice. Laws permitting these gag orders help protect against the risk of undermining an investigation by notifying the target. At the same time, gag orders increase the likelihood that illegitimate and unconstitutional surveillance will go unnoticed and unchallenged, since a target of an unlawful government surveillance order cannot challenge it unless they know it exists. Because broad authority to gag companies from notifying users of government demands for user data creates the potential for abuse, policymakers should consider whether existing legal authority permitting these gag orders is appropriately narrow. In particular, policymakers should consider whether the legal basis on which a gag order may be sought should be further limited, the duration of a gag order further restricted, or the ability to seek a gag order at all removed in certain circumstances. Policymakers should also consider whether companies should be permitted to make certain or additional aggregate information about government demands for user data publicly available, even if individual orders must be kept secret.[49]

Are gag orders on technology companies that receive government demands for user data constitutional?

Some tech companies have challenged the constitutionality of the gag order provisions for SCA orders and NSLs under the First and Fourth Amendments. Both the Third and Ninth Circuits have applied strict scrutiny to gag orders precluding providers from engaging in speech regarding requests for their customer's data and upheld the constitutionality of Section 2705(b) gag orders and NSL gag orders, respectively.[50] However, the Supreme Court has not addressed the constitutionality of these gag orders, and some advocates and commentators argue that they are prior restraints subject to an even higher level of scrutiny or that they do not satisfy

49 For example, policymakers should consider amending Section 604(a) of FISA to allow providers to report more granular statistical information about the number of demands they receive.

50 *Matter of Subpoena 2018R00776*, 947 F.3d 148, 155 (3d Cir. 2020); *In re National Sec. Letter*, 863 F.3d 1110, 1123 (9th Cir. 2017). The Second Circuit dismissed as moot constitutional challenges by Microsoft and Google to Section 2705(b) gag orders after disclosure was made to the affected customers. *Microsoft v. United States*, No. 20-1653 (2d Cir. May 14, 2021); *Google v. United States*, No. 19-1891 (2d Cir. May 14, 2021).

strict scrutiny.[51] In addition, although Congress enacted some limits on the duration of NSL gag orders as part of the USA FREEDOM Act in 2015, these limits are insufficient and do not cover all types of gag orders. In considering amendments to gag order provisions or new gag order provisions, policymakers should require gag orders to meet at least a strict scrutiny standard, *i.e.*, the gag order must be justified by facts showing that the order is narrowly tailored to promote a compelling state interest, and that there is no less restrictive alternative that furthers those aims. Moreover, to avoid Fourth Amendment concerns, authorization for gag orders should provide binding limits on their duration.[52]

When should companies notify users about legal demands for content removals or restrictions, and what information should be included in these notifications?

Notice from hosts of user-generated content that inform users when their content is removed or restricted based on a legal demand such as a court order gives users the information they need to legally challenge legal demands for content removals or restrictions or alert the public about the demands. In rare circumstances, it may be appropriate for hosts not to provide such notice: when they are prohibited from doing so by law, certain narrow emergency exceptions apply, or providing notice would be futile or ineffective.[53] When notice is provided, at minimum it should "identify the specific content that allegedly violates the law" and "inform the user that it was a legal takedown request."[54] Ideally, the notice should also include a copy of the legal order or other written demand, the identity of the government official, agency, or

---

51 *See, e.g.*, Al-Amyn Sumar, *Prior Restraints and Digital Surveillance: The Constitutionality of Gag Orders Issued Under the Stored Communications Act*, 20 Yale J.L. & Tech. 74 (2018); *Br. for Amici Curiae the Chamber of Commerce of the United States of America et al. in Support of Appellant*, *Microsoft Corp. v. United States*, No. 20-1653(L) (2d Cir. Dec. 21, 2020), ECF No. 125. Whether a particular gag order survives strict scrutiny may depend on the statutory authority under which it is authorized; for example, it may be easier for the government to meet strict scrutiny for nondisclosure under FISA than other laws.

52 *See* Br. for Amici Curiae the Chamber of Commerce of the United States et al., *supra* n.51. (arguing that the SCA's allowance for indefinite gag orders itself may give rise to a Fourth Amendment violation, and citing *Wilson v. Arkansas*, 514 U.S. 927, 930 (1995); *United States v. Freitas*, 800 F.2d 1451, 1456 (9th Cir. 1986); *United States v. Villegas*, 899 F.2d 1324, 1336-37 (2d Cir. 1990)).

53 Crocker et al., *supra* n.37 (explaining that emergency circumstances "should not be broader than the emergency exceptions provided in the Electronic Communications Privacy Act, 18 U.S.C. § 2702 (b)(8)" and that "[a]n example of a futile scenario would be if a user's account has been compromised or their mobile device stolen, and informing the 'user' would concurrently – or only – inform the attacker").

54 *Id.*

other entity who has made the legal demand and the legal basis for the demand. However, providing such detailed notice may be more expensive and time consuming for hosts, and may not be feasible for the smallest services.

There are additional considerations for hosts to weigh when government officials flag or refer content to the company, but the host removes the content under its own content policies. Clear notifications to users that the government was involved in flagging their content for review would allow users to bring legal challenges and draw public attention to this form of government action against their speech. However, such notifications may impose new costs on hosts, who may have to develop a process for tracking government referrals separately from other reports of violations of their content policies, so they can notify users of the government referrals. In addition, hosts may also object to providing user notices about government referrals because they fear it will give the false impression that a government referral *required* them to remove content pursuant to the host's content policy or improperly influenced their decision to remove content pursuant to their content policy. Such concerns can be mitigated by notices that clearly explain that a government official referred content for removal to the host under its content policy, and not under law, and that the host made the independent determination that the content at issue violated its content policy.

Notices about content removals and restrictions under a host's content policy but as a result of a government referral should include at least the same information as in user notifications about content moderation.[55] Ideally, the notice should also include a copy of the government referral and the identity of the government official or agency that made the referral.

What information should be included in user notifications about content moderation?

Notifications can enhance the legitimacy of content moderation by helping users understand why certain content is moderated. They can educate users about what content is allowed and forbidden on an intermediary's service, inculcating community values in users and helping users correct violative behavior. They can also shed light on content moderation decisions that are erroneous or with which users may disagree.

55 *See infra* User Notifications, Question 4.

To meet these goals, user notifications must contain enough information, communicated in a clear and understandable way, to actually inform users. More information is not always better; providing user notifications can be time and resource-intensive, and intermediaries must make decisions about the level of detail to include and how to design them to make them most effective. The information available may also depend on the type of service an intermediary offers and the content moderation methods it uses. The Santa Clara Principles, a set of principles for transparency and accountability in content moderation, recommend information that intermediaries' content policies and user notifications about content moderation decisions should include.[56] (While these recommendations provide a useful overview for policymakers of key considerations in the area of user notice, they are not model legislation and should not be incorporated wholesale into proposals that would mandate user notifications.)

Intermediaries and policymakers should also consider whether, in some instances, user notifications about content moderation may be counterproductive. For example, informing spammers about how and why their content has been moderated may enable them to evade moderation in the future. Similarly, users who intentionally violate an intermediary's content policies may respond to a notice that their content has been moderated or account has been banned or suspended by creating a new account through which they can continue to break the rules. While secret content moderation decisions may help prevent evasion of content policies, they can also undermine legitimacy, user education, and the ability to hold intermediaries accountable for their content moderation decisions.

Can user notifications about content policies and content moderation decisions be mandated in the United States, consistent with the Constitution?

As with transparency reports, American lawmakers considering mandates that require intermediaries to publish content policies and notify users of content moderation decisions should consider whether doing so is consistent with the First Amendment. In general, strict scrutiny applies to statutes that compel speech by private speakers. In addition, content policies and information about content moderation decisions go to the heart of intermediaries' exercise of editorial decisions about what content to allow on their services and how to display it, which is protected by the First

56 *Santa Clara Principles, supra* n.10.

Amendment. While requiring publication of content policies and user notifications of content moderation decisions may not be direct regulation of the editorial decisions intermediaries make, lawmakers should consider whether these requirements would exercise indirect governmental influence or control over intermediaries' editorial discretion and thereby violate the First Amendment.[57]

57 *Herbert v. Lando, supra* n.30; *Miami Herald v. Tornillo, supra* n.30.

Researcher Access to Data

Independent researchers, public policy advocates, and journalists seek access to data from hosts of user-generated content in order to investigate scientific or other academic questions, publish news or analysis, and inform advocacy and policy making. Improving researcher access to this data requires a common framework for understanding the current methods of access and the key questions – and the tradeoffs involved in their answers – that will shape policy decisions about regulating researcher access to this data.

////

Current Methods of Independent Researcher Access to Data

In general, independent researchers have three methods of obtaining access to data from hosts of user-generated content: (1) access to public data; (2) company-sanctioned access to public or nonpublic data; and (3) independent access to nonpublic data or data that is public but restricted.

Some data is available on the public internet.[58] Researchers collect this data manually or using automated methods such as scraping. For example, the website Pushshift[59] scrapes comments and posts from the social media website Reddit to create an archive of Reddit content that researchers have used to study issues such as social media echo chambers[60]

58 Whether online data is "public" may not always be immediately clear, and the definition of "public" may vary based on circumstances or statutory definitions.

59 Pushshift.io; Jason Baumgartner et al., *The Pushshift Reddit Dataset*, Assoc. for the Advancement of Artificial Intelligence (2020).

60 Matteo Cinelli et al., *The echo chamber effect on social media*, Proceedings of the Nat'l Academy of Sciences of the United States of America (Feb. 23, 2021).

or the effects of social networking deplatforming.[61] As discussed below, the scope of permissible scraping of public data is subject to ongoing policy and legal debate.

Some companies voluntarily make certain data available to researchers, often through Application Programming Interfaces (APIs).[62] APIs may be for general use or for use specifically by researchers. Companies may also voluntarily make data available through other datasets provided directly by the company or in partnership with a third party. Social Science One,[63] CrowdTangle[64] and the Twitter API for Academic Researchers[65] are all examples of company-sanctioned methods of researcher access to data. Company-sanctioned access may require researchers to apply to the company for access, satisfy criteria for access set by the company (such as affiliation with an academic institution), and obtain company approval of their research plans.

Finally, researchers use independent measures to gain access to hosts' data without company sanction, particularly from social networking companies.[66] The "data donation" method allows internet users to give their data directly to researchers, often using a custom web browser or browser extension installed by volunteers or paid participants.[67] The browser or extension collects and provides to researchers certain data from all of the internet sites that users visit or from particular social networks.[68] Researchers

61 Shiza Ali et al., *Understanding the Effect of Deplatforming on Social Networks*, Assoc. for Computing Machinery (2021).

62 APIs are "tools that allow programmers from outside the company to retrieve a set of data from company servers." Elizabeth Hansen Shapiro et al., *New Approaches to Platform Data Research*, NetGain Partnership at 13 (Feb. 2021).

63 *Social Science One* (last visited Nov. 29, 2021).

64 Will Bleakley, *About Us*, CrowdTangle (last visited Nov. 29, 2021). In April 2021, Facebook integrated CrowdTangle into its "integrity team," a move which some have criticized as intended to weaken the transparency provided by the tool in the face of negative information about Facebook reported as a result of CrowdTangle data.

65 *Twitter API: Academic Research Access*, Twitter (last visited Nov. 29, 2021).

66 This method is sometimes referred to as an "adversarial approach."

67 Giving users the ability to export their data, such as through interoperability services like Google Takeout, may also enable them to share historical data with researchers. *See* Ross James, *'What is Google Takeout?': How to use Google's simple tool for downloading all of your account data at once*, Insider (Jan. 23, 2020).

68 A browser extension is software that enhances the capabilities of a web browser, such as by allowing users to store passwords or block advertisements. Browser extensions used for data donation to researchers often copy specific content from the websites a user visits or a specific subset of those websites and transmits the data to the researcher. For example, the NYU Ad Observer browser extension copies the ads a user sees on Facebook or YouTube. *Ad Observer*, NYU Cybersecurity for Democracy (last visited Nov. 29, 2021).

use the collected data, often paired with demographic data from the participants, to examine how users encounter or interact with content and how social networks sites target content to users. For example, the MarkUp's Citizen Browser Project,[69] NYU Ad Observer,[70] and Mozilla Rally[71] all rely on data donation to gather social networking data.

Another method of independent access asks internet users to send data that may not be otherwise publicly available to a central platform or repository, which can then be accessed by researchers. For example, Junkipedia uses user submissions to create an annotated archive of mis- and disinformation from a range of platforms.[72] In a third method of independent access, researchers pose as users or advertisers to gather data. For example, researchers might pose as users by creating accounts with different demographic profiles or indicia to investigate patterns of bias[73] or as advertisers by placing ads on social media sites to investigate ad targeting.[74] Social media companies have resisted or shut down independent methods of data access in the past, such as when Facebook deactivated the accounts of two researchers from the NYU Ad Observatory, effectively blocking their research.

////

Enabling researcher access to data: Considering tradeoffs

Who should have access to data from hosts of user-generated content?

Because certain data can include highly sensitive and private information, restricting access to data to only particular entities and individuals is often desirable. Access could be restricted to certain categories such as "researchers" or "journalists." But defining these categories can be difficult and overly exclusive. For example, if "researchers" are defined as those with an academic affiliation,

69 *The Citizen Browser Project—Auditing the Algorithms of Disinformation*, Markup (Oct. 16, 2020).

70 Ad Observer, supra n.68.

71 *It's your data. Use it for a change.*, Mozilla Rally (last visited Nov. 29, 2021).

72 *About Junkipedia*, Junkipedia (last visited Nov. 29, 2021).

73 *See, e.g.,* Benjamin G. Edelman et al., *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment* (September 16, 2016). American Economic Journal: Applied Economics 9, no. 2 (April 2017) 1-22, Harvard Business School NOM Unit Working Paper No. 16-069; Sam Levin, *Airbnb blocked discrimination researcher over multiple accounts*, Guardian (Nov. 17, 2016); Kalhan Rosenblatt, *Senator's office posed as a girl on fake Instagram account to study app's effect*, NBC News (Sept. 30, 2021).

74 *See, e.g.*, Piotr Sapiezynski et al., Algorithms That "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences, arXiv (Dec. 16, 2019).

then journalists, civil society, independent analysts, government researchers, and 82% of all scientists and engineers[75] would be excluded from access. "Academic affiliation" would also have to be defined to determine whether, for example, affiliation with for-profit or foreign colleges and universities qualified.

Another approach would restrict access based on the intended use of the data. For example, access could be granted only to researchers whose research is in the public interest or meets other criteria intended to establish the research's importance or rigor, or only to researchers with a non-commercial purpose. Intended-use restrictions would require vetting the merits of proposed research or its non-commercial purpose and giving an entity or person (such as the company who holds the data, a government agency, or some other third party) the power to decide which researchers should be permitted to access data.

Vetting research to establish compliance with intended-use restrictions raises the risk of vesting too much power in the vetter to decide what research is in the public interest and what research is not; to lessen that risk, the vetter should be prohibited from discriminating based on viewpoint or the vetter's self interest. Even then, intended-use restrictions may still prohibit some worthy research; a non-commercial purpose restriction, for example, could inadvertently bar researchers who intend to sell books or news articles based on their research. However, given the privacy and other risks of granting researchers access to certain data held by hosts of user-generated content, screening research to determine whether it is in the public interest or meets other criteria may be appropriate.

Finally, access could be restricted based on an entity's or individual's ability to meet certain content-neutral criteria, such as the ability to conduct scientifically valid research (the meaning of which would have to be defined) and meet data security and privacy standards. Academic institutions that receive federal funding for research will typically have an Institutional Review Board (IRB) that could serve some of these functions, but the capacity of IRBs to conduct such assessments and enforce such standards is far from guaranteed.[76]

---

75 *S&E Workers in the Economy*, Nat'l Ctr. for Science and Eng'g Statistics (last visited Nov. 29, 2021).

76 *See* Simon N. Whitney, *Institutional review boards: A flawed system of risk management*, 12(4) Research Ethics 182 (2016); Prosperi, M., Bian, J. *Is it time to rethink institutional review boards for the era of big data?*, Nat. Mach. Intell. 1, 260 (2019).

What types of data do researchers seek access to, and why?

Different researchers seek access to different kinds of data to answer questions in fields such as the social sciences and computer science. Data from hosts of user-generated content can be broken down into a variety of categories.[77] One analysis has divided such data into three categories: (1) content data, such as posts or comments made by social media users or advertisements; (2) moderation data, or data about hosts' content policies and their decisions about enforcement of those policies; and (3) distribution data, or data about how and why users see particular content, including content recommendation algorithms.[78] Researchers may also seek access to other data, such as demographic information about users (which can provide important context to other categories of data), social networks or social graphs data, *i.e.*, data that shows how users of a social network are connected to each other, and other metadata. The data that researchers seek access to may be historical data or real-time data.

Different kinds of data raise greater or lesser privacy concerns, even within categories.[79] For example, content data about public social media posts may raise few privacy concerns, while content data about direct messages between users of a messaging service may be highly sensitive and protected from disclosure by law. Real-time content data about elections advertising may present different research opportunities, and raise different speech and privacy concerns, from historical data about ad targeting during a past election.

What online services should make data available to researchers?

While many hosts of user-generated content may have data that would inform research, most focus has been on access to data from consumer-facing online companies such as social media platforms. Defining what entities qualify as a "social media platform," however, is not always straightforward, since they may include social

77 Access to data unrelated to user speech or access to information, such as data about the finances or employees of hosts of user-generated content, customer data stored by cloud services, or government data held by companies with government contracts are outside the scope of this overview.

78 *See* Shapiro et al., *supra* n.62 at 17-24.

79 In addition, companies may be legally prohibited from sharing certain data, *see, e.g.*, 18 U.S.C. § 2702(a) (prohibiting a person or entity providing an electronic communication service to the public from knowingly divulging to any person or entity the contents of a communication while in electronic storage by that service, with limited exceptions) or may lose certain legal protections for data, such as those for trade secrets, if they disclose it publicly.

networking sites and applications, messaging services, content aggregation services, or even comment sections on news websites. Some of these services may have data that is more or less useful to research in the public interest and more or less sensitive than others. In addition, it may be necessary to draw distinctions in and between what data or how much data should be shared with researchers based on the size of the host to ensure that smaller hosts are not burdened by costs and obligations that may drive them from the market. These distinctions can be based on factors such as the age of the company, number of employees, revenues, or consumer usage, with upsides and downsides to each metric.[80]

How do we safeguard individual privacy while enabling broader access to data by researchers?

Company-held data can expose individuals' personally identifiable information, patterns of their online behavior, and the inferences that companies make about them. Certain data may be so sensitive that researchers should not be granted access to it at all, or should be granted access to it only for certain research projects. As a threshold matter, companies, lawmakers, and others considering the issue of researcher access to data should consider what data, if any, is so sensitive that it cannot be provided to researchers in some or all instances.

To the extent that researchers are granted access to personal or other sensitive data, companies, policymakers, and others must consider what privacy and data security protections to put in place. Privacy protections may be applied to the entirety of a research projector or in a multistage process. For example, a researcher could be granted access to an anonymized dataset for their research project, or they could be granted access to an anonymized dataset for their initial research and then later granted access to more sensitive data if they can demonstrate that their research is fruitful and access to additional data is necessary.

Privacy and data security can be protected through technical measures, access controls, legal liability, or a combination of methods. Common technical means of enforcing privacy include data aggregation, by which raw data is combined in a summary form, and differential privacy, which uses mathematical techniques to allow analysis of data while protecting its identifiable

80 Goldman & Miers, *supra* n.26.

characteristics.[81] These methods may require significant expertise and expense to implement and may limit the type of research that can be done. Access controls help protect user privacy by allowing researchers to access data only within environments where hosts can limit the analyses that researchers can perform, prohibit the copying or removal of data, and have in place data security measures such as encryption. This method may significantly constrain the type of research and the type of researchers who are able to conduct research, and it may prevent the sharing of data with research partners at other institutions, or other researchers who may seek to replicate a particular study. Finally, privacy can be protected through imposing legal liability for misuse of data in ways that violate privacy or security requirements, whether through generally applicable law that extends to certain data use, a statute written specifically to govern researcher access to data, or terms of service. Such methods, however, are only as effective as the enforcement mechanism and resources that accompany them.

How can companies and lawmakers eliminate unnecessary legal barriers to researchers' independent access to data?

Researchers that use independent methods to access data in the United States may face civil or criminal barriers to their work that lawmakers could eliminate or ameliorate. For example, changes or updates to the Computer Fraud and Abuse Act (CFAA) or Digital Millennium Copyright Act (DMCA) may remove or lessen the risk of liability for researchers.[82] In addition, voluntary carve-outs in companies' terms of service to permit research would remove the risk of civil liability for researchers who break terms of service by, for example, offering browser extensions that facilitate data donation. Congress could also require such carve-outs or immunize from civil liability researchers who break a companies' terms of service.

However, the CFAA, DMCA, and company terms of service can be important tools for limiting misuse of company data. As a result, companies and lawmakers should consider limiting any such carve-outs to apply only to research in the public interest. One challenge in this approach is how to write provisions that precisely distinguish between "white hat" or research in the public interest that should not be prohibited and other activity that in the guise of "research"

81 Bennett Cyphers, Understanding differential privacy and why it matters for digital rights, Access Now (Oct. 25, 2017).

82 Joseph Lorenzo Hall & Stan Adams, *Taking the Pulse of Hacking: A Risk Basis for Security Research* (Mar. 2018).

involves invasions of privacy, infringement of intellectual property, or other misuses that should be prohibited. In addition, the tradeoffs involved in intended-use restrictions on researcher access to data discussed above, such as the potential for abuse in vesting the power to decide what research is in the public interest in companies or government, apply here as well.[83]

Finally, in some instances, companies have used legal provisions or government consent decrees as a pretext for blocking researchers' access to data they hold on privacy grounds.[84] New federal privacy legislation or future government settlements with companies that violate existing privacy laws could state explicitly that research in the public interest or research that complies with particular criteria intended to protect user privacy are not forbidden on privacy grounds, to prevent companies' use of privacy laws or consent decrees as a basis for blocking independent methods of researcher access to data. Again, however, defining research in the public interest presents challenges.

Should researchers' access to data directly from companies continue to be at companies' discretion or be mandated in certain circumstances?

Current company-sanctioned methods of researcher access to data are voluntary. Voluntary provision of data to researchers allows a company and researchers to develop and experiment with different processes for providing access, which may lead to the development of new and innovative data-sharing methods. It also allows a company to decide what and how much data to share based on information that only the company may possess, such as the specific privacy needs of its users and the company's financial and other capacity to provide researchers with access.

However, company-sanctioned methods also allow companies to control which researchers can access their data, which may allow them to select researchers they perceive as being sympathetic to their interests or with whom they have previous relationships, potentially excluding researchers from less well-known or well-connected institutions. Some critics also argue that company-sanctioned methods of access give companies too much control over what data they will make available, for what purposes, and for

83 *See* Researcher Access to Data, Question 1, *supra*.

84 *See, e.g.*, Issie Lapowsky, *The FTC hits back at Facebook after it shut down NYU research*, Protocol (Aug. 5, 2021).

how long. In addition, purely voluntary company-sanctioned access raises the possibility that a company will intentionally manipulate data[85] or release erroneous datasets.[86]

Accordingly, some researchers, advocates, and lawmakers have proposed creating legal incentives[87] or even requiring companies to provide data to researchers. In choosing between incentives and mandates, lawmakers should consider that the First Amendment may prohibit the government from requiring hosts to provide certain moderation data and distribution data to researchers because doing so could violate their right to exercise editorial discretion over the user-generated content they host.[88] Incentivizing or mandating researcher access to data will also require policymakers to resolve all of the prior questions raised in this section: Who should have access to the data? What data should be provided? From what companies? And what privacy protections should be in place?

What is the best mechanism for providing researchers access to data from companies?

Company-sanctioned access to data – whether voluntary or in response to mandates or incentives – can occur through several possible methods, including:

- Making data directly available to researchers;
- Contributing data to a repository administered by a government entity; and
- Contributing data to a repository administered by a third party, such as an academic institution, existing non-profit, or new entity established for this purpose.

There are pros and cons to each of these methods. Directly sharing data with researchers allows use of existing mechanisms and infrastructure for access, such as APIs. However, this approach may be more burdensome for researchers and limit cross-company comparisons. Also, if the data is put in the hands of researchers, it may present privacy and security risks, such as researchers abusing their access by sharing data or inadequately protecting against leaks or other exposure of the data.

85 Hubert Horan, *Uber's "Academic Research" Program: How to Use Famous Economists to Spread Corporate Narratives*, Promarket (Dec. 5, 2019).

86 Craig Timberg, *Facebook made big mistake in data it provided to researchers, undermining academic work*, Wash. Post (Sept. 10, 2021).

87 Incentives could include offering companies protection from liability for privacy violations that result from the sharing of data with researchers.

88 *Herbert v. Lando*, *supra* n.30; *Miami Herald v. Tornillo*, *supra* n.30.

Creating a repository administered by either a government entity or third-party would potentially allow for standardization in data formats, methods of access, and privacy controls (while creating additional burdens and costs on companies to standardize data); however, it could create concerns about data security since the repository would be an attractive target for malicious actors seeking to gain unauthorized access to the data. A third-party repository could remove some of the self-interest involved if companies themselves are vetting researcher access, though it would need to be carefully designed to ensure that the third-party administrator was independent from companies that contribute data. In determining whether a repository administered by the government or a third-party is preferable, companies, policymakers, and others should consider whether it is preferable to have the government or a third-party in charge of vetting researchers. A repository administered by the government will also raise concerns about government surveillance of users, particularly if government access to the repository is not strictly limited.

Analysis, Assessments, and Audits

A key mechanism of tech company accountability is analysis of the company's business practices for their effects on individuals or their compliance with specific criteria. This analysis can take a variety of forms. For example, risk assessments are forward-looking, focused on the risks that a company's products or services pose and how the company can mitigate those risks. Audits, in contrast, are generally backwards-looking and focused on evaluating whether the company has met an objective set of standards or criteria. Both assessments and audits may be conducted internally or by independent third parties.[89] A primary goal of an audit is to provide the auditor's assurance that a company is meeting a particular standard. This does not always involve furnishing a detailed public report; in many cases, the auditor's opinion that the organization is in compliance with the audit criteria provides sufficient assurance. However, if a public report is published following an assessment or audit, it can offer some transparency about how a technology company operates and its impacts on the speech and privacy rights of users and communities. Third-party assessments or audits, in particular, can be important mechanisms for holding companies accountable to their commitments.

////

89 While the subject matter and approach of assessments and audits can vary widely, for purposes of this framework we address assessments and audits concerning companies' effects on user speech, access to information, privacy from government surveillance, and issues such as bias in the delivery of advertisements. Our use of "audits" to describe some of these evaluations is a more general usage than the specific processes of corporate financial and regulatory compliance audits; we note that there is a robust field of privacy audits, as a component of data protection regulation, that is out of the scope of this Framework.

Current Assessments and Audits

Some technology companies engage in risk assessments that they make available to the public. For example, Human Rights Impact Assessments (HRIAs) are an increasingly popular, though still rare, form of risk assessment focused on the impact of a technology company's practices and services on human rights.[90] The UN Guiding Principles on Business and Human Rights provide a set of guidelines for States and companies to prevent and address human rights abuses committed in business operations, which includes the expectation that companies will carry out human rights due diligence.[91] HRIAs are "a systematic approach to due diligence" through which a company examines "how its products, services, and business practices affect the freedom of expression and privacy of its users."[92] Companies may publish an annual human rights report[93] or discrete HRIAs on particular topics, such as a new or existing product or service[94] or their operation in particular countries.[95] The proposed Article 26 of the Digital Services Act in Europe would require certain ICT companies to engage in yearly risk assessments that consider certain specified risks, including their services' impact on particular human rights.[96]

Third parties also conduct analyses or assessments, either independently or in cooperation with the technology company, of whether company practice meets a set of pre-defined standards

90 Other stakeholders in the technology field also publish HRIAs; for example, the Global Internet Forum to Counter Terrorism, an NGO founded by technology companies to increase collaboration on online counterrorism efforts, recently published its first HRIA, BSR, *Human Rights Assessment: Global Internet Forum to Counter Terrorism*, BSR.org (2021), following advocacy from a coalition of human rights organizations. *See* Ctr. Democracy & Tech., *Human Rights NGOs in Coalition Letter to GIFCT* (July 30, 2020).

91 UN Working Grp. on Bus. & Human Rights, *The UN Guiding Principles On Business And Human Rights: An Introduction*, Office of the High Commissioner for Human Rights (last visited Nov. 30, 2021); BSR, *Conducting an Effective Human Rights Impact Assessment*, BSR.org (Mar. 2013). The UN B-Tech Project continues this important work, providing additional guidance on conducting human rights due diligence in the tech sector. *B-Tech foundational paper | Identifying human rights risks related to end-use*, Bus. & Human Rights Resource Ctr. (Dec. 14, 2020).

92 *2020 Indicators*, Ranking Digital Rights (last visited Nov. 30, 2021).

93 *See, e.g.*, *Corporate Social Responsibility*, Microsoft (last visited Nov. 30, 2021) (linking to the Microsoft Annual Human Rights Report).

94 BSR, *Google Celebrity Recognition API Human Rights Assessment | Executive Summary*, BSR.org (Oct. 2019).

95 *See, e.g.*, BSR, *Human Rights Assessment: Facebook in Myanmar*, Facebook (Oct. 2018); Chloe Poynton, *Our Assessment of Facebook's Human Rights Impacts in Sri Lanka & Indonesia*, Article One (May 12, 2020). As in these examples, companies often work with third-parties to conduct HRIAs.

96 *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, at Art. 26, COM (2020) 825 final (Dec. 15, 2020) [hereinafter "Digital Services Act"].

or criteria for responsible business practices, and publish these analyses or assessments publicly. Prominent examples include:

- Company Assessments by the Global Network Initiative (GNI), through which GNI independently assesses member companies on their progress in implementing the GNI Global Principles on Freedom of Expression and Privacy[97] with improvement over time, using a confidential review of companies' "systems, policies, and procedures" and responses to case studies.[98] GNI publishes a summary of each cycle's assessment process but the detailed reports remain confidential to the GNI Board;[99]

- The Ranking Digital Rights Corporate Accountability Index, an annual "evaluat[ion of] 26 of the world's most powerful digital platforms and telecommunications companies on their disclosed policies and practices affecting people's rights to freedom of expression and privacy," based on dozens of indicators in three main categories: governance, freedom of expression and information, and privacy;[100] and

- The Facebook Oversight Board, an independent body founded by Facebook to review Facebook and Instagram's content moderation decisions and issue policy advisory opinions on the company's content policies, which operates in a quasi-judicial style by reviewing individual cases against Facebook's values and community guidelines as well as international human rights standards, and publishing its decisions.[101]

Finally, independent third parties may also conduct and publish audits of technology companies, which are the systematic and independent collection and evaluation of objective evidence to determine whether specified audit criteria are fulfilled.[102] Technology companies may be covered by a variety of formal auditing

97 *The GNI Principles*, Global Network Initiative (last visited Nov. 30, 2021).

98 *Company Assessments*, Global Network Initiative (last visited Nov. 30, 2021).

99 *The GNI Principles at Work: Public Report on the Third Cycle of Independent Assessments of GNI Company Members 2018/2019*, Global Network Initiative (last visited Nov. 30, 2021).

100 *The 2020 RDR Index*, Ranking Digital Rights (last visited Nov. 30, 2021).

101 *Governance*, Oversight Bd. (last visited Nov. 30, 2021).

102 *See ISO 19011:2018(en) Guidelines for auditing management systems* at 3.1, International Organization for Standardization (last visited Nov. 30, 2021) (defining "audit").

requirements, including financial audits, privacy audits, and other evaluations of their compliance with particular regulations; often, these types of audits are not made available to the public and therefore do not serve a material transparency purpose.[103] In the past few years, however, several companies have also submitted to voluntary audits of their company practices based on concerns over systemic bias in the company's products, internal policies, or organizational structure.[104] These audits are often commissioned by a company, but they are conducted by independent third parties, such as a law firm or professional auditing firm. For example, in 2020, civil rights and civil liberties leader Laura W. Murphy and the law firm Relman Colfax PLLC published a final report on their Facebook Civil Rights Audit, which Facebook commissioned at the request of the civil rights community.[105] The field of civil rights auditing in the U.S. is nascent and the standards and practices for such audits are still in development.[106] The proposed Article 28 of the EU Digital Services Act would require certain very large online services to undergo formal yearly audits evaluating their compliance with various requirements in the Act; Article 33 would require companies to publish these audit reports along with a report on their implementation of any recommendations in the audit report.[107]

////

Improving Analysis, Assessments, and Audits: Considering Tradeoffs

Who should conduct analysis, assessments, and audits, and what criteria should independent assessors and auditors be required to meet?

Self-assessments allow companies to draw on their expertise and familiarity with their services to provide an evaluation that

103 *See, e.g.*, Michelle De Mooy, *How to Strengthen the FTC Privacy & Security Consent Decrees*, Ctr. for Democracy & Tech. (Apr. 12, 2018) (explaining that FTC privacy assessments of technology companies are not readily publicly available); *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1* at Art. 35 (requiring Data Protection Impact Assessments).

104 Laura W. Murphy, *Airbnb's Work to Fight Discrimination and Build Inclusion*, Airbnb (Sept. 8, 2016); *Three Year Review — Airbnb's Work to Fight Discrimination and Build Inclusion*, Airbnb (Sept. 10, 2019).

105 *Facebook's Civil Rights Audit – Final Report*, Facebook (July 8, 2020). Several chapters of the report addressed free expression issues such as content moderation and one chapter explicitly addressed privacy.

106 Laura W. Murphy, *The Rationale for and Key Elements of a Business Civil Rights Audit*, Leadership Conference on Civil & Human Rights (2021).

107 Digital Services Act, *supra* n.96 at Art. 28.

may be more holistic than that by an outside assessor or auditor. Self-assessments may also be significantly less expensive and more achievable for smaller and newer companies. However, self-assessments raise concerns about bias, *i.e.*, whether a company is objectively and impartially evaluating the effects of services on individuals' speech and privacy or the potentially discriminatory impact of their systems, and whether they have the cultural competency or other expertise to do so.

Third-party analysis, assessments, and audits may lessen concerns about bias, but only if the auditors and assessors are truly independent and are perceived as independent; assessors and auditors also need to have the requisite cultural competence and expertise. Accordingly, any voluntary or mandatory regime of third-party assessments and audits should establish requirements of independence and competency. Requirements for independence could include financial independence from the company being assessed or audited and elimination of other potential conflicts of interest, such as familial or business relationships between the assessor or auditor and the company. Important qualifications of assessors or auditors to consider are whether they have sufficient professional experience with and knowledge of technology companies and human rights, including free expression and privacy, as well as familiarity with the specific cultural context(s) in which the technologies are being used.

If assessments or audits are legally required, it may be necessary to establish a formal accreditation mechanism for assessors or auditors. Other forms of auditing may be helpful references for requirements or accreditation processes for assessors and auditors, such as international standards governing Environmental, Social, or Governance audits[108] or the International Organization for Standardization's requirements for accreditation bodies accrediting conformity assessment bodies.[109] These models may prove especially useful as the nascent assessments and audits of technology companies with respect to their business practices concerning speech, privacy from government surveillance, access to information, and other human rights are further developed.

---

108 *See* *ESG reporting and attestation: A roadmap for audit practitioners*, Association of International Certified Professional Accountants & Center for Audit Quality (Feb. 2021).

109 *ISO/IEC 17011:2017 Conformity assessment — Requirements for accreditation bodies accrediting conformity assessment bodies*, International Organization for Standardization (last visited Nov. 29, 2021).

What services should be assessed or audited and what are the appropriate assessment or audit procedures and criteria?

Different technology companies offer different services, and assessment and audit methods that may be appropriate for some services may not work for others. For example, an assessment or audit to evaluate the risks to speech and privacy caused by a social networking platform's use of algorithms in content moderation will need to examine different data from an assessment or audit of the risks to speech and privacy posed by a search engine sharing data with advertisers or government. In addition, assessment and audits most commonly evaluate technology companies against established criteria, such as international human rights standards, regulatory requirements, or voluntary principles to which a company has previously committed. Accordingly, any call to increase the number or scope of assessments and audits, either voluntarily or through legal requirements, must also consider the precise services that should be assessed or audited, the procedures to be used, and the criteria a company will be evaluated against.

What information from assessments and audits should be made publicly available?

Assessments and audits can provide valuable and valid assurances of company compliance with established criteria or standards based solely on the opinion offered by the individual or entity conducting the evaluation, if the evaluator is sufficiently credible. When evaluators publish not only their final opinion but also information about how they reached their conclusion, they can also enhance the transparency of technology company practices. Some kinds of analysis, like the Ranking Digital Rights evaluations of company practice, are conducted on the basis of already-public information. But not all information obtained in the course of an assessment or audit can be published. Assessors and auditors may need access to sensitive or confidential information from companies in order to create an accurate and complete evaluation, and companies may be willing to reveal this information only if it will not be publicly disclosed.

Companies may also seek to review reports before they are published in order evaluate whether any information they contain is privileged or protected by trade secret, and to redact this information or otherwise modify the report. If assessments or audits are to serve the additional purpose of transparency, however, final reports must reveal enough information to allow the public to

understand and evaluate them and hold companies responsible for the results, while not disclosing trade secrets or other proprietary information. Some assessments, like the GNI Company Assessments, try to strike this balance by providing information in anonymized or aggregate format. The competing interests in transparency and nondisclosure must be weighed against each other in determining what information and level of detail a final assessment or audit report should include.

Conclusion

Each form of transparency discussed in this framework – transparency reports, user notifications, access to data held by private companies for independent researchers, and public-facing analyses, assessments, and audits of technology company practices – holds unique promise and poses unique challenges. When considered together, they form a framework for greater transparency of tech company practices that impact user speech and privacy.

But transparency is not an end in and of itself – the purpose of tech company transparency with respect to user speech and privacy from government surveillance is to help us understand and check the ways in which these companies wield power and affect people's human rights. These various forms of transparency can empower individuals by providing them with useful, actionable information about the risks and benefits they face while using online services. Transparency can enable groups and communities to self-govern in online spaces, and can provide crucial information to researchers and journalists that help develop public understanding of our information environment. Transparency can also help inform the legislators and regulators charged with protecting individuals' rights.

In some instances, lawmakers may be able to respond legislatively to shape and direct tech company practices, such as when they engage in privacy abuses or anticompetitive behavior. As an example, in Europe, the proposed Digital Services Act seeks to address a wide range of issues in order to improve digital spaces and protect users' fundamental rights, including through increased transparency. However,

in other cases, lawmakers may not be able to legislate, even when companies make decisions that broadly impact society. For example, the growing call for companies that host user-generated speech to remove so-called "lawful but awful" speech – such as violent content, disinformation, some forms of hate speech, and harassment and threats that do not meet the relevant standards for violating the law. But international human rights law protections, and national constitutional standards, for free expression limit governments' ability to compel intermediaries to restrict lawful content. Companies may nevertheless act to restrict this kind of content – and in many cases, their users, and the broader community, want them to.

In these situations, transparency empowers those outside of government to act as watchdogs of tech companies. Transparency is key to the ability of nongovernmental organizations, the media, academics, and members of the public to bring public pressure to bear on technology companies. The knowledge that transparency creates allows civil society to influence these companies by critiquing, criticizing, and even shaming them. It allows users to make more informed choices about what services to use and which to boycott. Transparency also enables companies to learn from each other, to develop best practices and shared understandings of the challenges they face, and to apply the lessons learned from others' mistakes. Given the necessary constraints on government's powers to restrict speech, a multistakeholder approach, fueled by transparency, is a critical way to hold technology companies accountable and foster a robust environment for the fulfillment of our human rights online.

cdt.org

cdt.org/contact

Center for Democracy &
Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY