Transparency of Content Moderation Measures

Input paper for Workshop 1 – EU-US Tech and Trade Council Working Group 5
11 February 2022

This workshop will focus on current transparency reporting practices by digital platforms with respect to the content moderation decisions they take. What is the value of such public reporting practices? What data is currently being shared publicly? What are the limitations of these current practices? And what improvements to those practices are necessary?

1. The purpose behind voluntary transparency reporting mechanisms

Public disclosures about content moderation programs and enforcement procedures and transparency reports are generally "aimed at the general public".[1] Companies have created public transparency reports voluntarily from 2010 onwards in response to calls from academics, civil society organizations, and journalists to increase society's understanding of the amount and scope of requests for content takedowns from both individuals and governments.[2] In a later stage, governments have made similar calls. In 2018 the European Commission encouraged hosting platforms "to publish at regular intervals, preferably at least annually, reports on their activities relating to the removal and the disabling of content considered to be illegal content" (..) in order to better assess the effectiveness of notice-and-action mechanisms (..) and to ensure accountability there should be transparency vis-a-vis the general public".[3]

Civil society organisations have called for the creation of transparency reports to ensure that companies' enforcement of their own terms of service would be more "fair, unbiased, proportional and respectful of users' rights".[4] In this sense, transparency reports can be seen as an "element of due process procedures"[5] and as such they were seen as a first step to hold companies publicly to account for actions they took or failed to take against specific types of content.

However, the usefulness of these voluntary practices has been questioned. Transparency scholars have warned how "opaque" forms of transparency can actually be used to obfuscate processes and practices.[6] Gorwa and Garton Ash have alluded that current voluntary transparency reporting practices are less a tool towards accountability then they are a tool for companies "to regain the trust of the public, politicians and regulatory authorities".[7] Similarly, Wagner et al have argued that "many actors prefer to create the illusion of transparency rather than actually engaging in transparent practices"[8] given that transparency reports aren't currently linked to a broader framework that can hold platforms to account for their content moderation decisions, or lack thereof. In general, the data that is made available describes in incomplete terms the content moderation process that a platform engages in. It generally describes a certain type of inputs (e.g. notices received) and outputs (e.g volumes of content taken down), and sometimes goes one step further to show the volumes of complaints received against these decisions (see section 2 below).

[1] Mark MacCarthy, *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry*. Transatlantic Working Group, 12 February 2020, https://www.ivir.nl/publicaties/download/Transparency_MacCarthy_Feb_2020.pdf
[2] Jack Goldsmith and Tim Wu, *Who Controls the Internet?: Illusions of a Borderless World.* New York: Oxford University Press, 2006.
[3] European Commission Recommendation (EU_ 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018H0334&from=EL
[4] The Santa Clara Principles on Transparency and accountability in content moderation, https://santaclaraprinciples.org/
[5] MacCarthy, id at 1.
[6] Robert Gorwa and Timothy Garton Ash, Democratic Transparency in the Platform Society, in Social Media and Democracy: The State of the Field (Cambridge, 2020), edited by Nate Persily and Joshua Tucker.
[7] Idem.
[8] Ben Wagner et al, Regulating transparency?: Facebook, Twitter and the German Network Enforcement Act. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. January 2020, 261–271, https://doi.org/10.1145/3351095.3372856

However, these reporting practices do not describe the scale and severity of harms that users experience on the platform, nor do they give any insights into the scale or impacts of these harms. They also do not give insights into a platforms' operations, nor a sense of performance assessment for their content moderation processes.

Additionally, these voluntary mechanisms are not equipped to shed light on potential violations of consumer protection laws by the platforms, nor do they provide sufficient details to scrutinize how crucial features of these companies work in practice. In that sense they reflect the inherent limits of public disclosure mechanisms, which should strive to balance the often-contradictory goals of improving transparency, protecting privacy rights, and the rights of companies. Voluntary public disclosure mechanisms are also limited by a company's significant incentives to avoid disclosing unflattering information. As such, current voluntary transparency reports don't provide a sufficiently nuanced understanding of how companies actually formulate, apply and enforce their Terms of Service.

At the same time this does suggest a need for a tiered approach to disclosure, whereby regulators, auditors, and scientific researchers should have access to more granular forms of data in order to perform their respective functions. Different mechanisms can be envisaged to allow regulators and scientific researchers to get access to this data, which will be the subject of Workshop 3. Nevertheless, given the important role some companies play, partly due to their reach, in facilitating public debate and contributing to information ecosystems in societies around the world, it is important to achieve a more granular level of public transparency as well.

2. Disclosure of metrics and categories of content

The Open Technology Institute maintains a living a living document [9] that highlights which metrics and categories of content Facebook, Instagram, Reddit, TikTok, Twitter, and YouTube currently report on in their transparency reports related to the enforcement of content rules. As the Trust and Safety Professionals Association points out[10], most reports would now include aggregated metrics such as:

- How many pieces of content, and associated accounts, were removed or otherwise enforced against for violating a company's policies?
- How many pieces of content were reported by users (regardless of whether the content was removed)?
- How many users reported content that was later removed?
- How much violating content was removed before a user ever reported it?
- How long was violating content available on the product before it was removed?
- How many users have asked the company to reconsider a moderation decision?
- How often was the original moderation decision overturned after a second review?

In recent years, platform transparency reports have increased in scope and depth on a broader range of topics, including, for example, advertisements or activities directed against coordinated inauthentic behavior. Partly in response to growing public pressure, more data types are becoming publicly available which use different methods to measure user exposure to content that violates a company's Terms of Service. These include the "prevalence" metric[11] released by Facebook, the "violating view rate"[12] released by YouTube, and the "reach of policy-violating Pins"[13] metric released by Pinterest, all of which use different methods to measure user exposure to violating content.

[9] Spandana Singh, Leila Doty, The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules, https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/
[10] Trust and Safety Professional Association, Transparency Report Categories, https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/transparency-report-categories/
[11] https://about.fb.com/news/2019/05/measuring-prevalence/
[12] https://blog.youtube/inside-youtube/building-greater-transparency-and-accountability/
[13] https://policy.pinterest.com/en/transparency-report

There are existing multi-stakeholder efforts particularly around Terrorist Use of the Internet, such as the ongoing OECD Terrorist and Violent Extremist Content (TVEC) Voluntary Transparency Framework effort[14], as well as a transparency reporting template by Tech Against Terrorism[15], focusing on smaller companies' capabilities to report. The Santa Clara Principles on Transparency and Accountability Around Content Moderation, and associated implementation toolkits for advocates, companies, and regulators is another example.

3. The limitations of public transparency reports

In those instances where there is no legal obligation for companies to disclose specific metrics (see section 4), companies have the sole discretion to decide which metrics they report on, how they calculate the data they share with the public, and which metrics they do not report on. For example, the recent disclosures by Facebook whistleblower Frances Haugen revealed that Facebook only removes between 3-5 percent of hate speech on its platform.[16] In its Community Standards Enforcement Report 2021, however, Facebook reports proactively removing between 80-90 percent of hate speech.[17] Both figures seem to be correct but based on different metrics. Facebook has an interest in publicly sharing data that creates a more favorable perception of the company. Without Haugen's disclosures, the public may not have had an alternate means by which to assess the validity of Facebook's own hate speech removal figure.

These different classifications make meaningful comparisons between companies "extremely difficult".[18] For instance, Keller and Leerssen point out that "reports that track how many notices a company received cannot fruitfully be compared to reports tracking how many items of content they were asked to remove, since one notice may list any number of items".[19]

Scholars have noted that current voluntary disclosure practices may obscure more than they reveal. As Sing and Doty point out "some platforms continue to lump categories of content together in a manner that obscures potentially valuable insights. For example, YouTube reports on its moderation of spam and misleading content together. While there may be technical challenges to collecting and segmenting this data on the backend, combining them prevents users and researchers from understanding how the company is tackling misleading content specifically".[20] The Integrity Institute has detailed which additional data could be made available publicly in order to "understand the scale and cause of harms occurring on social media platforms", and "enable the public to validate that social media companies are using best practices in responsibly designing and building their platforms".[21]

Moreover, opaque content moderation practices such as Facebook's "CrossCheck" whitelist, or so-called "three strike policies" are entirely hidden from the public eye. There is to date also no effective way for independent outsiders to determine the effectiveness of automated content moderation systems, including their impact on freedom of expression, in particular from the perspective of marginalized communities and non-English languages.[22] Keller has similarly pointed out the need for additional transparency mechanisms, stating that "aggregate data in transparency reports ultimately just tell us

[14] https://www.oecd.org/digital/transparency-reporting-on-terrorist-and-violent-extremist-content-online-8af4ab29-en.htm
[15] https://transparency.techagainstterrorism.org/
[16] Whistleblower aid - Re: supplemental disclosure of securities law violations by Facebook. Inc, available at https://drive.google.com/file/d/1CZrCqyCHJ7L1EHPeorGBoQpKhKFRrMC5/view
[17] https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/
[18] Wagner, ibid note 8.
[19] Daphne Keller, Paddy Leerssen, Facts and where to find them: empirical research on internet platforms and content moderation. In: Nate Persily and Joshua Tucker, Social Media and Democracy: The State of the Field and Prospects for Reform (Cambridge University Press, 2020).
[20] Sing and Doty, idem note 9.
[21] Integrity Institute, Metrics and Transparency, 22 September 2021, https://static1.squarespace.com/static/614cbb3258c5c87026497577/t/617834d31bcf2c5ac4c07494/1635267795944/Metrics+and+Transparency+-+Summary+%28EXTERNAL%29.pdf
[22] OSCE, Spotlight on artificial intelligence and freedom of expression. OSCE, 2021, https://www.osce.org/files/f/documents/8/f/510332_0.pdf

what platforms themselves think is going on. To understand what mistakes they make, or what biases they may exhibit, independent researchers need to see the actual content involved in takedown decisions".[23]

4. Legislative response to gaps in public transparency reporting: ongoing proposals in the EU and the US to establish baseline metrics for platforms to report on

This criticism has given rise to questions about whether legislation can and/or should establish baseline metrics platforms can report on, and to what extent different types of transparency and accountability mechanisms need to be explored.[5] I

In Germany, the German Network Enforcement Act (NetzDG) specifies that platforms that receive more than 100 notifications about unlawful content per year must publish a public transparency report in German every 6 months. As Wagner et al point out[24], these requirements include a

- general outline of how criminal activity on the platform is dealt with
- a description of the (content moderation) mechanisms in place
- the number of the complaints
- organisational and human resources (dedicated to content moderation)
- membership of industry bodies
- number of complaints for which an external body was consulted
- number of complaints that were deleted
- time span of deletion or blocking procedure in place and
- measures to inform the user who submitted the complaint, as well as the users whose content is under investigation.

In the EU, the Terrorist Content Regulation[25] specifies that transparency reports should include

- information about the hosting service provider's measures in relation to the identification and removal of or disabling of access to terrorist content;
- information about the hosting service provider's measures to address the reappearance online of material which has previously been removed or to which access has been disabled because it was considered to be terrorist content, in particular where automated tools have been used;
- the number of items of terrorist content removed or to which access has been disabled following removal orders or specific measures, and the number of removal orders where the content has not been removed or access to which has not been disabled (..) together with the grounds therefor;
- the number and the outcome of complaints handled by the hosting service provider
- the number and the outcome of administrative or judicial review proceedings brought by the hosting service provider;
- the number of cases in which the hosting service provider was required to reinstate content or access thereto as a result of administrative or judicial review proceedings;
- the number of cases in which the hosting service provider reinstated content or access thereto following a complaint by the content provider.

[23] Daphne Keller, Some humility about transparency, http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency

[24] Wagner et al

[25] Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, article 7, available at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32021R0784

Both in the EU and the US, legislators have proposed legislation that would further address some of these issues.

For example, the proposal for a Digital Services Act would require providers of intermediary services to produce transparency reports on a number of prescribed topics including:

> (a) the number of takedown requests,
>
> (b) the number of notices it received, categorized by content type and originator,
>
> (c) the number and type of measures taken that affect the availability, visibility and accessibility of information provided by the recipients of the service and the recipients' ability to provide information, categorised by the type of reason and basis for taking those measures; and
>
> (d) the number of complaints received about its content moderation measures, the basis for those complaints, decisions taken with respect to those complaints, the average time needed for taking those decisions and the number of instances where those decisions were reversed.

Additionally, depending on their size and societal impact, platforms would need to make data available about any use of automatic means for the purpose of content moderation, including specification of the precise purposes, indicators of the accuracy of the automated means in fulfilling those purposes and any safeguards applied.

In the United States, bipartisan bills such as the PACT Act, introduced by Senators Thune and Schatz, would require that platforms of a certain size publish quarterly transparency reports which would include:

(a) the total number of instances in which illegal content, illegal activity, or potentially policy-violating content was flagged (i) due to a user complaint; or (ii) internally, by (I) an employee or contractor of the provider; or (II) an internal automated detection tool;

(B) the number of instances in which the interactive computer service provider took action with respect to illegal content, illegal activity, or known potentially policy-violating content due to its nature as illegal content, illegal activity, or known potentially policy-violating content, including content removal, content demonetization, content deprioritization, appending content with an assessment, account suspension, account removal, or any other action taken in accordance with the acceptable use policy of the provider, categorized by (i) the category of rule violated; (ii) the source of the flag, including government, user, internal automated detection tool, coordination with other interactive computer service providers, or personnel employed or contracted for by the provider; (iii) the country of the information content provider; and (iv) coordinated campaign, if applicable;

(C) (i) the number of instances in which an information content provider appealed the decision to remove potentially policy-violating content; and (ii) the percentage of appeals described in clause (i) that resulted in the restoration of content; and

(D) a description of each tool, practice, action, or technique used in enforcing the acceptable use policy.

At the sub-national level, recent efforts to discuss content moderation practices and policies have been introduced into U.S. state legislatures, including California Bill AB-587 would require platforms to disclose content moderation practices. Importantly, both bills put transparency reporting in a wider accountability framework. This is important, since governments run the risk of legitimizing, rather than constraining, companies' ineffective moderation practices if they simply mandate companies to deliver unverifiable transparency reports.

5. Key questions for the workshop

1. What is the purpose of public transparency reporting as a complement to other forms of transparency towards privileged third parties such as regulators, auditors, journalists, or scientific researchers?

2. Which metrics can/should platforms be reporting on publicly?

    a. What are the most useful metrics to share to achieve the purposes in question 1?
    b. How frequent should these metrics be shared, and why?
    c. Are there lessons that can be learned from existing multistakeholder efforts around content moderation on specific issues like Terrorist Use of the Internet?

3. What incentives can governments create to disclose meaningful transparency reports and encourage deeper discussion on moderation measures with the public?
    a. What role can/should transparency mandates play to improve current practices? What legal barriers exist to achieve transparency mandates? How could the proportionality of transparency mandates be ensured to avoid unintended consequences on platforms?
    b. What role can/should voluntary agreements, codes of conduct, or standardization efforts play (with independent party review/auditing or rigorous certification programs)?