EU-US Trade and Technology Council Working Group 5
Workshop 1: Transparency of Content Moderation Measures

Friday 11 February 2022

Participants:

· Alexandre de Streel (University de Namur)
· Angel Diaz (UCLA)
· Daniela Stockmann (Hertie School)
· Daphne Keller (Stanford University)
· Gerard de Graaf (European Commission)
· Jeff Allen (Integrity Institute)
· Karen Kornbluh (GMF)
· Marie Therese Sekwenz (Sustainable Computing Lab Vienna)
· Mark MacCarthy (Georgetown University)
· Martin Husovec (LSE)
· Mathias Vermeulen (AWO)
· Nahema Marchal (UZH)
· Nick Diakopoulos (Northwestern University)
· Paul Barrett (NYU)
· Peter Harrell (US Government)
· Prabhat Agarwal (European Commission)
· Robert Gorwa (WZB Berlin)
· Tim Wu (US Government)

Summary

Introduction

The meeting began with introductory remarks from the representatives of the European Commission and the US Government, highlighting that transparency of content moderation practices was one of the most likely areas of convergence between the EU and the US when contemplating initiatives to meaningfully cooperate on issues of data governance and technology platforms.

Both sides stressed the importance of embracing a tiered approach to disclosure practices, whereby regulators, auditors, and scientific researchers could have access to different types of data in order to perform their respective functions compared to the types of information that should be disclosed to the user or the general public.

The purpose of public transparency reporting

The discussion began with a historical view, noting that what we understand as content moderation transparency began in earnest in the 2010s in the context of government takedown requests and government access requests. The objectives of public transparency reporting practices has evolved since then, with many new objectives raised during the discussion: improving the public's understanding of platforms' content moderation practices, better protecting users from unfair treatment, encouraging companies to spend more resources on these reports and allowing a greater diversity of perspectives and critiques into content moderation practices. Crucially, public transparency measures were seen by some as meaningless if they don't operate within a broader accountability framework. Participants generally agreed that current content moderation transparency practices don't go far enough to meet these goals.

Some participants argued that there is no overall logic behind what companies decide to disclose publicly. A lot of disclosed aggregated data sheds light on governmental requests to moderate content for instance, whereas there is very little data available that would explain to what extent users of a platform have been exposed to content that goes against a platform's terms and conditions.

Social scientists argued that public facing transparency measures currently help reveal trends that can generate hypotheses about content moderation practices. A single set of information can yield different hypotheses. For instance, data showing low levels of user complaints against a platform's content moderation decisions can either be a sign that users identified few instances of problematic content, or that they actually do not believe that a mechanism to submit complaints works.

It was argued that current content disclosure regimes as mandated by the German NetzDG law or the EU Code of Conduct on Hate Speech currently fall short to allow researchers to properly test hypotheses as well. Researchers stated they are not able to actually understand platforms' content moderation practices on the basis of these reports. Being able to determine what is actually happening on a platform is only possible with deeper access to data, which could include access to privacy-sensitive information or proprietary information. Yet, there was an agreement that limiting access to transparency data exclusively to academics is not desirable from an accountability perspective: besides the disclosure of aggregated figures and anonymized datasets, it is in the

public interest to force platforms to publish properly redacted reports that would be based on more privileged forms of access to transparency data.

Current transparency reporting practices

Participants mentioned the following useful metrics that platforms could make available first, to researchers or auditors, and secondly, in redacted form to the general public:

- Data on the lifecycle of harmful content (e.g. who uploaded it? How was it uploaded? Does the uploader have a track record of previous violations? How was the content distributed? When/where was content seen by other users?)
- Information on what platforms do with certain pieces of content (how does it get exposed to users? How was it ranked? Does the platform show it to users that didn't choose to follow this type of content?). These metrics can help understand why users were exposed to harmful content, rather than just how many times they were exposed.
- Content-based datasets (e.g. samples of impressions on platforms) can also be useful to understand the effectiveness of platforms' content policies, and would enable users to highlight to platforms the content that they haven't taken down but should have). Samples of content impressions are useful because they provide a view into 'super spreader' content (content that was seen by a larger group of users). One participant nonetheless noted that providing this information can be expensive for smaller platforms given the different languages and cultural nuances that must be captured in different geographies.
- Time-related information (e.g. how much the spread of content has changed over time).
- Information on changes of user behaviour (e.g. did users alter their behaviour as a result of particular interventions?)
- Contextual information (e.g. breakdown per geography or provenance)
- Historical data archives/repositories that clearly document how and when content was taken down or were subject to other actions (e.g. labelling).
- Information on the spread of content across different platforms to understand the effectiveness of cooperation between different platforms (e.g. which platform found the content in question first? Was the content uploaded again to another platform?)

The importance of making these metrics available ultimately depends on their purpose. Regulators might have an interest in these datasets to scrutinize the effectiveness of a company's content moderation practices, whereas for the general public it might be important to know some of these data points in order to understand the danger of being exposed to specific types of content.

In terms of the frequency of transparency disclosures, having platforms provide daily updates on their content moderation practices could be justified given that platforms are constantly evolving their moderation systems and testing new models.

Potential incentives to improve current reporting practices

Participants debated to what extent content moderation transparency requirements could be standardized in order to improve the comparability of transparency reports across different companies. It was argued that this would require a separate discussion about definitions of key terms that are currently being used differently by platforms. For example, should transparency reports list the number of *notices* received in particular categories, or the number of *items* reported?

Others argued that standardization of public transparency reports can not only be useful to assess the effectiveness and performance of specific companies' content moderation practices, but also more fundamentally might allow the creation of a marketplace where consumers can make an informed choice about which platform they decide to use. One participant made an analogy with car safety testing procedures, where the existence of uniform safety tests allows consumers to pick a car that aligns the most with their preferences.

Since the Digital Services Act's voluntary standardisation provisions do not have a jurisdictional limit, and will require some engineering efforts anyway, it was argued that emerging EU rules on content moderation could result in platforms globally applying at least some public transparency requirements.

Finally, participants debated the opportunities and challenges of governments mandating specific types of transparency reporting. Participants agreed that companies would never disclose some data points unless forced to by law. Similarly, providing access to raw data, content (removal) or privacy sensitive data would need to be mandated by law, and fit within a broader auditing or accountability framework. The obligations on platforms would need to be proportionate to their size. In the US context, participants seemed to agree that mandating transparency directly was better than tying transparency requirements to a broader reform of S230 CDA.

One of the challenges with having enforcers decide whether transparency obligations have been properly implemented is that platforms might be incentivised to not disclose information that could make them liable to enforcement. In order to avoid that risk one participant mentioned the idea of creating a separate entity that could verify the veracity of the transparency data that a platform discloses. Another option to deal with this problem is to deal with content moderation transparency in the same way as financial reporting. This means that platforms would need to certify the claims that they make about their products. Whereas auditors would have access to the full company reports, one area of best practice to explore is to make redacted versions of these reports available to the public as well.

Participants also noted the need to separate transparency mandates from governmental requests for platform data, as regulators' power to gain access to platform information could be abused by authoritarian governments. It will therefore be important for content moderation regulators to be independent from governments in the same way as the Federal Trade Commission (FTC) and Federal Communications Commission (FCC) are formally independent from the US Government.

Final remarks

Representatives from the European Commission and US Government closed the meeting by noting that continuous dialogue with researchers is important in order to arrive at a better understanding of what meaningful public transparency could look like, and how to create incentives for companies to deliver the best reporting mechanism. Both sides look forward to using these workshops as inspiration to agree on joint principles that both the EU and the US could potentially endorse in the future.