TOWARDS MEANINGFUL AND INTEROPERABLE TRANSPARENCY FOR DIGITAL PLATFORMS

2022 Outcome of the UN IGF Coalition on Platform Responsibility

Luca Belli, Yasmin Curzi, Clara Almeida, Natália Couto, Roxana Radu, Rolf H. Weber, and Ian Brown

This document was elaborated through a participatory process organized via the mailing list of the IGF Coalition on Platform Responsibility and led by the Coalition's Coordinators. After having collected initial inputs, a draft was shared using a collaborative pad to receive suggestions and comments. The draft was also presented during a meeting of the Action Coalition on Meaningful Transparency to expand the spectrum of stakeholders providing feedback. The Consolidated version of this Statement has been presented at the IGF 2022 session of the Coalition.

Abstract: Transparency and accountability are crucial mechanisms to ensure that decisions from private and public organizations are legitimate and trustworthy. Concerning specifically digital platforms, scholars, experts and digital rights organizations have been developing studies on what such concepts mean and highlighting their relevance for policymaking. Historically, platforms are self-regulated actors, making their own private ordering with content moderation and terms of service. Pressures towards transparency, accountability and democratic commitments have been leading policymakers worldwide to produce legislation in order to address such deficits. Nevertheless, national regulation can often lead to more internet fragmentation. In this article, we develop the idea of "interoperable transparency" to address this issue, suggesting standardized practices and methodologies for its implementation.

Table of contents

1. Context: Why meaningful and interoperable transparency?

The digitalization of societies and economies, turbocharged by the recent pandemic, has led digital platforms – notably, social networks – to acquire a crucial role as intermediaries of public discourses and, increasingly, any kind of private or public service. The limited regulation that until very recently framed these platform activities – or the absence of it in most jurisdictions – has allowed these intermediaries to become private regulators of these activities.

These actors define content moderation techniques and regulate data collection and usage, primarily based on self-regulatory tools – such as Terms of Service, Community Policies, Privacy Policies, etc. – and the very same technical architecture of the platforms (Belli, 2022; Belli et al., 2017; MacKinnon, 2013).

In this context, one of the few points of agreement amongst many interested stakeholders is that transparency and accountability are vital for increasing users' trust, assuring platforms' responsibilities and the legitimacy (Haggart & Keller, 2021) of their actions -– i.e., that they are based on the rule of law and due process.

Nevertheless, transparency and accountability can be abstract concepts that need more robust definitions and practical guidelines to become meaningful (Vogus & Llansó, 2018; Weber, 2021a), especially when private companies are the only actors mediating their activities and practices through self-reports.

Another point of agreement is that transparency and accountability should not merely imply the disclosure of information but also the auditability of the disclosed information (Ausloos & Leerssen, 2020). To ensure that content moderation procedures and the data about the spread of disinformation, hate speech, and other undesirable content can be verified, it is also essential to ensure the accuracy of what is, in fact, the information provided by large online platforms (Wagner & Kuklis, 2021).

Notably, the activities carried out by large platforms regarding the moderation of content can be deemed as remarkably similar to State traditional powers and functions as these players can decide which speech can be maintained and promoted and which should be removed or deprioritized – a *quasi-judicial* power – according to the rules developed by them – a *quasi-legislative* power (Chenou and Radu, 2017). They also enforce their decisions with content moderation and sanctions for users who do not comply with them – a *quasi-executive* power (Belli, 2022; Belli et al., 2017; Belli & Venturini, 2016).

The private decisions taken by large platforms, which are carried out without democratic legitimacy (Haggart & Keller, 2021), may affect an enormous number of individuals, companies, and even public organs. Their terms and policies are unilaterally drawn up and considered by the literature as adhesion contracts, in which users have no power to bargain (De Filippi & Belli, 2012; Prausnitz, 1937; Radin, 2012) or capacity to *input* (Haggart & Keller, 2021). Moreover, the guidelines for content moderation activities, including filtering, flagging content or taking it down, combine artificial intelligence tools,

manual reviews, and reports from users in ways that remain opaque to the user (Radu, 2019).

Platforms' architectures (Lessig, 2006) are another aspect of concern regarding transparency, particularly regarding content moderation and free speech. There are significant informational asymmetry issues between platform providers and users, mainly due to algorithmic opacity (Pasquale, 2016). Algorithms are silently responsible for organizing platforms' news feeds, governing the informational flows, and organize advertisements.verts. Such activities – which expose users to specific types of harmful content, political ads and propaganda, misinformation and disinformation – can considerably impact users' safety and the well-functioning of democracies.

For example, Facebook's emotional contagion study (Albergotti & Dwoskin, 2014) has revealed that algorithmic content recommendation could significantly affect users' emotional stability. Furthermore, the way in which algorithms prioritize or downgrade content – typically on social media "timelines" – can considerably impact users' access to information and civil society's capacity for mobilization and exercise of the right of peaceful assembly (Tufekci, 2015). By the same token, large platforms play new roles in the contestation of democratic processes and the protection of human rights online, as the cases of Cambridge Analytica and online hate speech inflaming the genocide in Myanmar (Human Rights Council, 2018) show.

In such context, researchers and whistleblowers have repeatedly evoked the necessity of improving platforms' transparency and accountability to protect the full enjoyment of human rights (Ananny & Crawford, 2018), offering different approaches to achieve such goals. Suzor *et al*. (2019), for example, originally suggested the idea of meaningful transparency to foster change in the transparency approaches of major platforms. In this perspective, to comply with transparency requirements promoted by civil society organizations, such as the criteria set by the Santa Clara Principles, platforms started to voluntarily present reports containing general data about content removal, user security, data protection, and compliance with national provisions regarding copyright violations, terrorism, child abuse, and crime prevention.

Several scholars have elaborated an increasingly large body of research exploring the issue. For instance, Rieder and Hofmann (2020) coined the concept of observability to refer to transparency as a procedure. In this sense, the authors define true transparency as the capacity of civil society and academia to scrutinize platforms' actions. Based on Seaver's (2017) ethnographical investigation of companies' practices regarding the significance of "algorithms" provides valuable insight into the difficulty – even for internal developers – of identifying and assessing "the algorithms" and their impact.

Wagner and Kuklis (2021) caution that the lack of verifiable data on platform content moderation practices that public institutions could audit concretely leads to a situation where it is simply not possible for public regulators to know how disinformation works in practice.

Thus, regulators should recognize that even with existing audits, disclosures, and justification mechanisms (Rieder & Hofmann, 2020, p. 5), the promise of transparency might not be fulfilled. Critically, the concept of

observability has been converging with meaningful transparency, for it tries to call out the necessity of a framework or regulation encompassing continuous and constant observation of platforms' activities.

Moreover, it is essential to note that even opening platform APIs to researchers (e.g., the Twitter Developers' Program or Crowdtangle for Facebook) might present limitations as they are typically based on self-regulation and, therefore, platform policy changes might easily limit or revoke researchers' access to information, at least as long as not a co-regulatory framework is implemented. Such an approach has merits because, according to Ausloos & Leerssen (2020), providing data access for independent researchers is also a matter of public interest since they can aid in diagnosing harms, developing and enforcing evidence-based policies, and mobilizing accountability.

This scenario has led researchers and civil society advocates to argue that interoperability can offer a solution not only to competition issues in the digital ecosystems (Brown & Marsden, 2013; Ausloos & Leerssen, 2020), but also to counter platform practices that undermine freedom of expression due to a lack of transparency. For example, some authors propose that platforms' infrastructures could be interoperable by compelling social media platforms to share APIs or "middleware" acting as common content-curation services ascribing users control over the information they see on various platforms (Docquir & Stasi, 2020; Keller, 2021).

Other authors propose to establish a generative interoperability structure building online public and civic spaces (Tarkowski et al., 2022). Such measures could mitigate platforms' economic power concentration in some countries, foster competition, enable users' data portability rights, and increase user freedom of information. Another possibility is that companies could substantially design interoperable rules and content moderation practices and share their methodologies for content moderation and reports' development.

2. Characteristics: Key elements of meaningful and interoperable transparency

Criticism regarding digital platforms' lack of transparency in their activities has increased to the point that the meaning of the terms has become empty (Gillespie, 2019, p. 212). However, as Suzor *et al*. (2019) explain, demands for greater transparency usually refer, explicitly or implicitly, to greater disclosure of information that would – supposedly – lead to greater accountability and trust in that institution, as also argued by Rieder and Hofmann's (2020) conceptualization mentioned above of transparency as the means to scrutinize platforms' decisions.

If we consider "meaningful transparency" as the disclosure of information in a way that can lead to proper accountability, then transparency should be treated as a relative concept depending on the target of responsibility. Concretely, to make transparency meaningful, any regulatory policy should clarify the object of transparency, its audience, why disclosing the information is essential, and its goals. In addition, more information is not always better, i.e. transparency should not be measured in terms of quantity but in terms of quality (Weber

2021b, pp. 78/79). To whom and for what reason must digital platforms be transparent?

Scholars like Ananny and Crawford (2018, p. 985) have also questioned the relationship between the proposal to increase transparency and the desired result regarding the possibility of accountability. Leerssen (2020), in dialogue with the abovementioned authors, explains that the first question to be asked is "whom do transparency measures serve?" in the logic of "targeted transparency." Then, by examining the targeted audience (to whom the companies entailed the transparency measures), it is possible to design accountability propositions for the platforms in question (Leerssen, 2020, p. 19).

The information can target three groups of stakeholders, as classified by Leerssen (2020): (i) individual users of the platforms, to inform them about how (personal) information will be used and organized by the platform and about removal decisions of content or account that may occur; (ii) regulatory bodies, public supervisors, and other auditing bodies; and (iii) civil society, the general public, and independent researchers.

As the recipients of the information disclosed by the platforms might be diverse, the transparency criteria might also vary. Nevertheless, platforms often only provide general and heterogeneous information on content exclusion in their reports, avoiding extensive and standardized explanations about content moderation policies such as content recommendations (Cobbe & Singh, 2019) and the application of other rules aimed at regulating online expression (Goldman, 2021). Information on recommendation activities – e.g., the functioning of algorithms recommending, suppressing, pricing, prioritizing, or downgrading specific types of data – is vital for policymakers, regulators, or researchers that could better address platforms biases and propose solutions.

In the case of public interest functions performed by social media platforms, the use of automated decision-making tools for enforcing and balancing users' fundamental rights requires further scrutiny (Marsden and Meyer, 2019; de Gregorio and Radu, 2022). It also is essential for users and content creators willing to understand the content remedies applied to them. Indeed, research indicates that they might even change potential harmful behavior with a better understanding of how platform rules are applied in a given community (Jhaver et al., 2019).

In the EU regulatory context, the Digital Services Act (Regulation (EU) 2022/2065) provides several starting points for transparency in content moderation (as well as recommender systems) both regarding terms and conditions, and data access for researchers (Schwemer, 2022).

In this sense, we could define three different moments in the moderation decision-making process that require the provision of information by the platforms:

I - the establishment of moderation policies and rules;

II - the decision-making of moderation actions against a specific user's content or account;

III - enforcement of policies and rules of moderation in the social network environment, including as regards the existence of remedies.

3. Conditions: A framework suggestion

To improve the current platforms' transparency scenario, content moderation policies and frameworks should effectively contribute to making social media platforms accountable for harmful activities while offering precise indications regarding how to be held accountable. We understand that for achieving meaningful transparency, companies should outline measures that aim at (1) enabling observability, including by increasing the accuracy of data about transparency measures on platforms' content moderation, and (2) creating interoperable standards on transparency.

Figure 1 - Model Framework for Meaningful and Interoperable Transparency for Digital Platforms

1. Reports with context	To make sure that quantitative data can be assessed from a qualitative perspective, platforms should make available datasets, including qualitative information on:	Which content was reported; Which measures were taken by the platform; Which procedure was adopted by them (maintenance, removal, de-prioritization, etc.); To what extent due process requirements were applied; What was the consequence of the user appeal (changed the original decision or not).
	Platforms should also provide support for independent institutions and researchers (Ausloos & Leerssen, 2020, pp. 83–84), encouraging risk assessment, audits, investigations, and other types of reports on platforms' practices and activities.	
2. Standardized and shared rules	From a substantial perspective, platforms should share detailed and intelligible information on the following:	Their content moderation rules; The employment and functioning of automated algorithmic moderation systems; Due process procedures;
	From a methodological perspective, platforms should:	Collectively standardize the presentation of reports in coordination with other stakeholders; Make data continuously available in an interoperable, understandable and machine-readable format and auditable by interested third parties; Publish their initiatives – including risk assessments reports – regarding the identification and prevention of biases in their algorithms and content moderation procedures; Allow the use of interoperable content moderation APIs.

First, at a systemic level, platforms should make continuously available datasets about content moderation practices that could be analyzed and independently audited by multiple stakeholders (regulators, researchers, civil society, journalists, etc.). Report obligations with numeric data are insufficient and fail to provide essential information to evaluate and contextualize platform practices.

When data is included, categories of interest for various stakeholders need to be presented in a disaggregated form. Providing information with the total amount of blocked accounts or removed contents may allow elaborating quantitative statistics but makes it impossible to have a qualitative analysis of the problem at stake in the lack of information on due process (i.e., the reasoning of the removal and users' appeal) or the functioning of content de-prioritization and shadowbans (Myers West, 2018).

Providing context is especially relevant for preventing biases in platforms' moderation activities, enabling false positives to be corrected, and would also help in understanding and mitigating asymmetries regarding platforms' measures in different countries.

As emphasized previously, another relevant aspect often left aside in platforms' reports is the functioning of content moderation algorithms. This dimension is essential to understand how platforms' recommendation systems operate and to increase corporate accountability when the recommendation of harmful content has nefarious consequences on users' (mental) health or democratic processes, especially when platforms earn hefty advertisement revenues deriving from the sharing of such harmful content.

In this sense, we propose that a **Model Framework for Meaningful and Interoperable Transparency for Digital Platforms** should at least encompass the aspects described in Figure 1.

4. References

Albergotti, R., & Dwoskin, E. (2014). Facebook Study Sparks Ethical Questions—WSJ. Wall Street Journal. https://www.wsj.com/articles/facebook-study-sparks-ethical-questions-1404172292

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Ausloos, J., & Leerssen, P. (2020). Operationalizing Research Access in Platform Governance What to learn from other industries? Algorithm Watch, 109.

Belli, L. (2022). Structural Power as a Critical Element of Social Media Platforms' Private Sovereignty. 22.

Belli, L., & Venturini, J. (2016). Private ordering and the rise of terms of service as cyber-regulation. Internet Policy Review, 5(4). https://doi.org/10.14763/2016.4.441

Belli, L., & Zingales, N. (2022). Interoperability to foster open digital ecosystems in the BRICS. Chinese Academy of Cyberspace Studies.

Belli, L., Zingales, N. (Eds.). (2017). Platform regulations: How platforms are regulated and how they regulate us: official outcome of the UN IGF Dynamic Coalition on Platform Responsibility (1st edition). FGV Direito Rio.

Brown, I., & Marsden, C. (2013) Regulating Code: Good Governance and Better Regulation in the Information Age. MIT Press.

Chenou, J.-M., & Radu, R. (2019). The "Right to Be Forgotten": Negotiating Public and Private Ordering in the European Union. Business & Society, 58(1), 74–102. https://doi.org/10.1177/0007650317717720

Cobbe, J., & Singh, J. (2019). Regulating Recommending: Motivations, Considerations, and Principles (SSRN Scholarly Paper No. 3371830). Social Science Research Network. https://doi.org/10.2139/ssrn.3371830

Docquir, P. & Stasi, M. (2020). The Decline of Media Diversity — and How We Can Save It. Centre for International Governance Innovation. https://www.cigionline.org/articles/decline-media-diversity-and-how-we-can-save-it/

De Filippi, P., & Belli, L. (2012). The Law of the Cloud V the Law of the Land: Challenges and Opportunities for Innovation (SSRN Scholarly Paper No. 2167382). https://papers.ssrn.com/abstract=2167382

De Gregorio, G., & Radu, R. (2022). Digital constitutionalism in the new era of Internet governance, International Journal of Law and Information Technology, 30(1), 68–87, https://doi.org/10.1093/ijlit/eaac004

Gillespie, T. (2019). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press. https://doi.org/10.12987/9780300235029

Goldman, E. (2021). Content Moderation Remedies. Michigan Technology Law Review, 28(1), 1–60.

Haggart, B., & Keller, C. I. (2021). Democratic legitimacy in global platform governance. Telecommunications Policy, 45(6), 102152. https://doi.org/10.1016/j.telpol.2021.102152

Human Rights Council (2018). Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar (12 September 2018). https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_CRP.2.pdf.

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–27. https://doi.org/10.1145/3359252

Keller, D. (2021). The Future of Platform Power: Making Middleware Work. Journal of Democracy, 32(3), 168–172. https://doi.org/10.1353/jod.2021.0043
Leerssen, P. (2020). The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems (SSRN Scholarly Paper No. 3544009). Social Science Research Network. https://doi.org/10.2139/ssrn.3544009
Lessig, L. (2006). Code: And Other Laws of Cyberspace, Version 2.0 (2nd Revised ed. edição). Basic Books.
MacKinnon, R. (2013). Consent of the Networked: The Worldwide Struggle For Internet Freedom (Reprint edition). Basic Books.
Marsden, C. & Meyer, T. (2019) Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism. European Parliamentary Research Service. https://research.monash.edu/en/publications/regulating-disinformation-with-artificial-intelligence-effects-of
Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366–4383. https://doi.org/10.1177/1461444818773059
Pasquale, P. of L. U. of M. F. (2016). The Black Box Society: The Secret Algorithms That Control Money and Information (Reprint edição). Harvard University Press.
Prausnitz, O. (1937). The Standardization of Commercial Contracts in English and Continental Law. Sweet & Maxwell, Limited.
Radin, B. A. (2012). Federal Management Reform in a World of Contradictions. Georgetown University Press.
Radu, R. (2019). Negotiating Internet Governance. Oxford: Oxford University Press.
Rieder, B., & Hofmann, J. (2020). Towards platform observability. Internet Policy Review, 9(4). https://doi.org/10.14763/2020.4.1535
Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. Big Data & Society, 4(2), 2053951717738104. https://doi.org/10.1177/2053951717738104
Schwemer, S.F. (2022). Digital Services Act. A Reform of the e-Commerce Directive and Much More. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4213014
Suzor, N. P., West, S. M., Quolding, A., & York, J. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. International Journal of Communication, 13, 1526–1543.
Tarkowski, A., Bloemen, S., Keller, P. & de Groot, T. (2022). Generative Interoperability. Building Online Public and Civil Spaces. https://openfuture.eu/wp-content/uploads/2022/03/InteroperabilityReport.pdf.
Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency Symposium Essays. Colorado Technology Law Journal, 13(2), 203–218.
Vogus, C., & Llansó, E. J. (2018). Making Transparency Meaningful: A Framework for Policymakers (No. 1; p. 45). Center for Democracy and Technology. https://cdt.org/wp-content/uploads/2021/12/12132021-CDT-Making-Transparency-Meaningful-A-Framework-for-Policymakers-final.pdf
Wagner, B., & Kuklis, L. (2021). Establishing Auditing Intermediaries to Verify Platform Data. In Regulating Big Tech: Policy Responses to Digital Dominance (pp. 169–179). Oxford University Press. https://doi.org/10.1093/oso/9780197616093.003.0010

Weber, R.H. (2021a). Internet governance at the Point of No Return, EIZ Publishing Zürich. https://eizpublishing.ch/wp-content/uploads/2021/05/Internet-Governance-at-the-Point-of-No-Return-V0_78_2-20210503-digital.pdf
Weber, R.H. (2021b). From Disclosure to Transparency in Consumer Law, in: Mathis K., Tor A. (eds.), Consumer Law and Economics, Springer, Cham, 73-87.