‘Before transparency’: The Digital Service Act and the legitimisation of platform power

in *(In)visible European Government: Critical Approaches to Transparency as an Ideal and a Practice*

(eds) Päivi Leino-Sandberg, Maarten Zbigniew Hillebrandt, Ida Koivisto

(Preliminary Draft do not circulate)

Marta Maroni, University of Helsinki

Contents

1. Introduction

Social media platforms are powerful actors in current society. They shape the visibility of content and services in a way that has proven disruptive to democratic arrangements. Caught in one scandal after the other, including their impact on electoral outcomes, the spread of disinformation, hate speech, and violation of fundamental rights such as freedom of expression, the EU is trying to intervene in the field of content moderation. Transparency is a much-celebrated solution to deal with the abuse of power of these companies. Since the inception of the internet, self-regulation has been the preferred model of regulation, and the Digital Service Act (DSA) makes it mandatory by imposing the proceduralisation of major digital companies. Self-regulation was enshrined in the E-commerce Directive (2000). This exempts platforms from liability in the case of illegal content disseminated by third-party users. However, the E-commerce directive did not specify how the ‘notice and take-down’ system, on which self-regulation is based, should have been implemented, thereby raising concerns about the removal of content. The DSA, while leaving the liability system untouched, imposes due diligence obligations on major digital platforms. Transparency requirements have become focal for the entire regulatory regime. This chapter discusses transparency as a practice and as an ideal in the context of platform governance and content moderation. Practically, it discusses transparency standards as adopted in the DSA, and theoretically, the chapter intervenes in the ongoing narrative that portrays the DSA as a major regulatory break in platform regulation. Drawing on critical transparency studies, this chapter unravels the ambivalent nature of transparency and argues that transparency measures function more as a

legitimising force for digital media platforms than as manoeuvres against the power structure of these platforms.

The chapter underscores the role transparency plays in making the power of these platforms acceptable and how, at the same time, it obscures more profound questions concerning the legitimacy of these actors in ordering contemporary society. This work is situated in the debate on digital constitutionalism and offers insights into how European regulation becomes co-constitutive of the power structure of digital media platforms. To be clear, although the chapter problematises the function of transparency and the project it operationalises, it does not argue for less transparency; rather, it subscribes to the necessity to control and regulate how platforms operate. Furthermore, it acknowledges that the final version of the DSA contains extensive transparency requirements.

The chapter is structured as follows. Part 2 stresses how issues of disinformation and hate speech are intertwined with the business model of digital platforms, which is based on the attention economy and introduces *the question* of transparency in content moderation. Part 3 discusses the general provisions of the DSA and contextualises why transparency occupies a central stage in the new EU regulatory approach. Moreover, it advances the argument that content moderation is still predominantly understood as a content takedown and that aspects of content distribution are less elaborate. This part also stresses the continuity between the freedom afforded to internet intermediaries for more than 20 years, and the need to increase the bureaucratisation of major digital media platforms. Part 4 analyses the kinds of transparency mechanisms required under the EU framework. It examines how transparency is mostly constructed as publicity, quantification and procedural fairness. Although the DSA foresees important solutions, they ultimately remain designed to protect the business models of these platforms. Drawing on science and technology studies and critical dataset studies, this part highlights how the provisions in the DSA do not fully consider the mediating power of technology (e.g. users experience interface, the governance of the Application Programming Interface) and the extent to which transparency measures could be a mediated device to manage visibility. Overall, the DSA takes a data-centric approach and leaves aside issues of algorithmic transparency. The combination of these scholarships enables us to raise questions about the efficacy of the transparency measures under the DSA and anticipate a possible problem with the EU strategy. Part 5 wraps up and concludes the chapter.

2. Transparency as a problem and as a solution in platform content moderation

> "*Facebook's closed design means it has no oversight—even from its own Oversight Board, which is as blind as the public*. Only Facebook knows how it personalizes your feed for you. It hides behind walls that keep the eyes of researchers and regulators from understanding the true dynamics of the system…"[1]

On 4 October 2022, whistle-blower Frances Haugen read aloud the words above as she testified before the United States Committee on Commerce, Science, and Transportation. In her statement, Haugen warned about the harm that Facebook's products cause: "They harm children, stoke division, and

[1] Statement of Frances Haugen, United States Senate Committee on Commerce, Science and Transportation Sub-Committee on Consumer Protection, Product Safety, and Data Security (October 4, 2021) p 3

weaken our democracy, and much more".[2] She firmly denounced how Facebook's own internal research confirms the problems Facebook itself causes: the amplification of division, extremism, and polarisation. Moreover, she showed how Facebook resolves conflicts between its profits and our safety by consistently favouring its profits. She continued:

> "Almost no one outside of Facebook knows what happens inside of Facebook. The company leadership keeps vital information from the public, the U.S. government, its shareholders, and governments around the world".[3]

She emphasises that any attempt to regulate Facebook will result in an ill-founded one if regulators do not have adequate knowledge of what to regulate:[4]

> "The core of the issue is that no one can understand Facebook´s destructive choices better than Facebook because Facebook gets to look under the hood. A critical starting point for effective regulation is transparency: full access to data for research not directed by Facebook. On this foundation, we can build sensible rules and standards to address consumer harms, illegal content, data protection anticompetitive practices, algorithmic systems, and more. As long as Facebook is operating in the dark, it is accountable to no one. And it will continue to make choices that go against the public good".[5]

Haugen described Facebook as "a company with control over our deepest thoughts"[6] and in calling upon states' regulation, she argued that Facebook could *be a better* social media platform with proper oversight. Although the lack of transparency in digital media platforms is a mounting problem, we may also notice its ambiguous nature. Although transparency promises oversight of how power operates, it also obscures the question of whether that power should be in place in the first place. After all, social media platforms run on a business model, which monetises people's ties and relations (Schwarz Ori, 2019), and poses great personal and social risks given the way it influences the consumption of information. In other words, transparency, as an ideal, comforts that something can be controlled while sidelining discussions as to whether digital platforms such as Facebook and Google should exert so much power in current society (Maroni, 2022). Digital media platforms are advertising spaces that match users with news, services, and applications, and they deliver personalised content. To be good at what they do, social media platforms need to keep users engaged with their services. This is what characterises an attention-economy business model: the more a user engages with the platform, the more the platform can personalise the form of addiction of its users (Bhargava & Velasquez, 2021). In other words, attention is a scarce resource that platforms need to maximally exploit (Bhargava & Velasquez, 2021; Wu, 2018).

---

[2] ibid. p 1
[3] ibid. p 2
[4] Statement of Frances Haugen (n_) and a similar point also noted by Daphne Keller and Paddy Leerssen, Keller D and Leerssen P, "Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation" in Nathaniel Persily and Joshua A Tucker (eds), Social Media and Democracy: The State of the Field, Prospects for Reform (Cambridge University Press 2020)
[5] ibid. p 2
[6] ibid. p 3

This business model entails digital platforms controlling the type of information delivered to users; platforms might decide to tailor the information according to the users' pre-existing interests to maintain their attention. Scholars have also highlighted a correlation between a system that commodifies our attention and false content that can exploit attentional bias (Lewandowsky & Pomerantsev, 2022). Lewandowsky and Pomerantsev pointed out that people are known to attend news that is "predominantly negative or awe-inspiring". The same is true for "messages couched in moral-emotional language" (Lewandowsky & Pomerantsev, 2022).

Platforms moderate their content based on their own internal rules, which concern the distribution of content as much as its removal. For each of these two processes, there are specific procedural and architectural issues to consider in order to gain a thorough picture of how content moderation works. Community standards—the policy face documents accessible on the platform's website—might represent an entry point into the governing rules of social media platforms, especially regarding content takedown. However, community standards do not explain or give an account of other internal decisions guiding the visibility and removal of content. Much information about what platforms do is protected by property rights. Leaked documents or whistle-blower revelations have relatively contributed to opening the system for outsiders and revealed how difficult it is to conceptualise what transparency should be about in the context of platform governance.

For example, regarding the distribution of content, the latest "Facebook Papers" (2021) revealed, among other things, how the Facebook algorithm differently weighted different users' reactions: the "angry" reaction was weighted five times more than the simple "likes. This formula fostered widespread problematic content, as emotionally charged reactions occurred more often in relation to toxic content. Users eventually started being exposed to less problematic content when Facebook set the weight on the angry reaction to zero.[7] This is one but an example of information that is not publicly disclosed and the implications of social media affordances, such as a "like" button or interface design, which determine how content appears online.

Content moderation is constituted by a set of norms and practices (also understood as sociotechnical design) adopted by social media outlets for both making content accessible and inaccessible. The first refers to the way platforms rank, prioritise, and optimise content, whereas by making content inaccessible, we refer to how platforms downgrade, obscure (shadow ban), block, and remove content and accounts. It is important to highlight this double aspect of content moderation activities and the different ways they shape access to content, as the primary concern has thus far settled on the removal of content or suspending account rather than the distribution of content, for example (Daphne Keller & Paddy Leerssen, 2020). This was not without a reason, as we shall see below, the legal framework has been mostly preoccupied with the problem of takedown (and account suspension) as a violation of freedom of expression rather than with the issue of the optimisation of content, which instead entails thinking in terms of media pluralism—that is, the quality and plurality of information. The way we frame the problem (mostly on content removal) has implications for the type of transparency standards we need to develop. Under the DSA, the conceptualisation of content moderation broadens, and in addition to interpreting content moderation, that is, content takedown and account suspension

[7] Cristiano Lima (2021) <https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/> accessed 8 February 2022.

(deplatformisation), Art. 3(t) also takes into account the problems of inaccessibility, such as demotion and demonetisation, and it mentions questions of media freedom and pluralism.

3. The European proceduralisation turn in content moderation

The Digital Service Act will enter into force on 17 February 2024. This regulation adds and harmonises new rules for digital media platforms and their content moderation practices. The DSA is advancing one of the most *up-to-date* regulatory frameworks for content moderation and is considered to contribute to the "Brussels Effect" (Bradford, 2019), that is, the development of new global regulatory standards for content moderation.[8] This new regulation is also situated within the debate on European constitutionalism and internet regulation (Guilherme Cintra Guimarães, 2019) and problematises the narrative that Europe is limiting platforms' powers (Pollicino & Gregorio, 2021) and breaking from the early laissez-faire. Although this is, to a certain extent, a valid statement, because the EU is imposing stricter rules on content moderation, it is plausible that that the need to establish a complex and decentralised apparatus to control what platforms do is part of the trajectory of the power previously afforded to these companies, and it also serves to maintain that power in place. In this context, transparency develops along the lines of the privatisation of the legal system (and its algorithmisation) and the need to empower the public to exert oversight as well as to promote "users' agency"; that is, users can modify the criteria for receiving advertisements.[9] Amidst these developments, transparency has become a technology that produces new forms of knowledge by quantifying and qualifying different aspects of content moderation (Sally Engle Merry, n.d.).

The purpose of this section is to provide the necessary background to understand the role that transparency plays. The DSA is a very complex piece of legislation that makes self-regulation mandatory for all big-tech platforms and hosting service providers. The DSA represents an essential step forward in content moderation, given the societal risks embedded in the design and function of platforms, as well as questions of ownership. Under the DSA, digital media platforms are requested to adopt due diligent obligations to deal with illegal content and tackle the risks associated with the use of their services. The DSA codifies some of the self-regulatory mechanisms already adopted by platforms such as Facebook and Twitter (e.g. flagging systems, automation to detect illegal content, internal redress mechanisms, transparency reports) and combines them with new extensive requirements and a complex system of oversight mechanisms.

Further, the DSA envisages voluntary solutions such as the Code of Conduct Illegal Speech & Code of Practice on Disinformation. It should be highlighted that these "voluntary solutions" aim to become regulatory standards for disinformation problems; this framing, to a certain extent, cast doubts on the

[8] Papaevangelou, C., 2022. Digital Services Act, Brussels Effect and the Future of the Internet - JOLT. [online] JOLT. Available at: <http://joltetn.eu/digital-services-act-brussels-effect-and-the-future-of-the-internet/> [Accessed 11 September 2022].

[9] In this regard, it seems that similarly to privacy and data protection, the contributes to constructs the idea of a rational person increasingly responsible for its own actions. (For an insightful perspective about the privatisation of a person see (Lindroos-Hovinheimo, 2021).

substantive grip of the regulation (EDMO, Implementation of the Code of Practice on Disinformation: Lessons from the assessments and proposals for the future European Digital Media Observatory). The EU Commission is also institutionalising the decision-making power of platforms. This is visible when it entrusts them with important decisions, such as the enforcement of fundamental rights or imposing obligations on very large online platforms and search engines to prevent abuse of their systems by taking risk-based actions. This results in leaving the definition of what is risk and how to balance it with fundamental rights to the discretion of the platform (vrf blog). To better understand how transparency is functional in maintaining the power of these platforms, the section below discusses what is at the core of the regulation, which is the exemption from liability and the new diligent obligations.

To start with, the DSA complements the E-commerce directive (ECD), which so far has been the European foundational regulation for liability exemption of internet intermediaries. Similar to the E-commerce directive, the DSA protects internet intermediaries' commercial activities *inter alia by listing the* cases when an intermediary cannot be held liable for the third-party information they transmit and store. The DSA builds on different categories of internet intermediaries contained in the ECD: mere conduit, caching[10] and hosting. *Online platforms* belong to the category of hosting services, but under the DSA, the difference between the two is that a hosting provider stores and disseminates information to the public. In other words, they make information available to a potentially unlimited number of third parties (Art. 2 DSA). Very large online platforms (VLOPs) refer to platforms that have a significant societal and economic impact, reaching at least 45 million users in the EU and establishing additional obligations. The DSA transposes ECD Section 4 (Art. 12-15) into Articles 3-5, which establishes a limitation of liability on all forms of illegal activities taking place on intermediary service providers. Similar to the ECD, the DSA maintains that intermediary service providers should not be liable for illegal third-party content when they do not initiate or have actual knowledge of the illegal activity or information and are unaware of the facts or circumstances concerning the illegal activity or information. To grant exemption from liability, the intermediary should act in a neutral and passive way, which means that the hosting service provider should lack knowledge of illegal activity. Furthermore, the hosting provider should act in a technical, automatic, and passive way (Recital 42, DSA). Scholars have criticised the binary construction of passive and active; it is unsuitable to conceptualise how online platforms moderate their content (Stalla-Bourdillon & Thorburn, 2020; Maroni Marta & Brogi Elda, 2021), as they can be both active and passive; for example, they optimise, rank, and filter content. As I have argued elsewhere, the European Court of Justice (hereinafter ECJ or the Court), in an attempt to square the circle about complex questions that have emerged in applying the ECD, maintained the active and passive construction in a way that preserved the construction of neutrality and internal market rationality and thus not profoundly challenging the power structures of digital media platforms. For example, the ECJ has clarified that *offering services for remuneration* does not impact exemptions from liability (*Google France SARL*, 2010). It ruled that neutrality applies to an actor who has confined itself to a merely technical process of data (*Google France SARL*, 2010). In

[10] According to Art 2 of the DSA, 'mere conduit' are service that consists of the transmission in a communication network of information provided by a recipient of the service, or the provision of access to a communication network, whereas a 'caching' service that consists of the transmission in a communication network of information provided by a recipient of the service, involving the automatic, intermediate and temporary storage of that information, for the sole purpose of making more efficient the information's onward transmission to other recipients upon their request;

this regard, the Court corroborated the idea that the automatic processing of information equates to a lack of knowledge, thereby maintaining the construction that platforms are neutral. Furthermore, the Court has ruled that the imposition of a filtering system for illegal content consists of general monitoring, which is prohibited under the ECD. However, it has also clarified that automation can be deployed to target content with equivalent meaning to one considered illegal. The DSA codifies the ECJ case law and also refers to thorny issues discussed under the European Court of Human Rights case law (e.g. Delfi Case), that is, how to consider commenting space of an online newspaper (DSA, Recital 13). The ECD provided intermediaries with immunity from liability where "upon obtaining actual knowledge of illegal activities, they acted expeditiously to remove or disable access to information concern". This is commonly called the 'notice and takedown' mechanism. However, the ECD did not specify how this should have been implemented, and to remedy this gap, the DSA introduced a notice *and action mechanism* (DSA, Art. 14), which stipulates that all providers of hosting services should put in place user-friendly mechanisms to notify illegal content. The notice should be reasoned; for example, it should explain the illegality of the content, contain information about the identification of the content, and include a good faith statement. The reception of the notice inevitably makes the hosting provider *aware* of the illegal content, which has to decide on its removal. This system creates both practical and conceptual tensions. On the one hand, the notice and takedown mechanism represents the most immediate and efficient solution, but it also results in private companies setting standards for freedom of expression. It is a given that major platforms need to take responsibility for what happens in their spaces, and in light of this, the DSA broadens the conditions for the exemption of liability by establishing a "Good Samaritan" clause. The clause clarifies that when internet intermediaries carry out *voluntary own initiative* investigations to detect, identify, and remove illegal content, they will not incur any sanction (DSA, Art. 6). This provision introduces a new exception to the liability framework (based on a lack of knowledge). Further, the "Good Samaritan" clause interfaces the second pillar of the liability framework that is the prohibition on Member States to impose *general monitoring obligations* on providers to monitor the information which they transmit or store. This second principle is considered particularly significant to protect freedom of expression and privacy online and to protect the right to conduct business with internet intermediaries (cf. DSA, p 12).

At this juncture, we may notice how the prohibition against imposing *general* monitoring obligations regards states, but monitoring does not include platform activities, for example, profiling (which is a form of monitoring, but it is personal, hence no general). There is a very neoliberal conception of state regulation being a threat to society and the private, bringing progress. Thus, the legal construction of neutrality, based on lack of knowledge, has been problematic, given that profiling is about to know about the transmitted information (since the system works on previous preferences), and it is also what allows social media platforms to curate and rank and personalise information with consequent social risks associated with them. However, domestic authorities can always issue an injunction upon a hosting service to stop and prevent infringement. Over time, the ECJ has differentiated between *general monitoring* and *specific monitoring*—that is, monitoring applicable *to a specific case* and *limited* in terms of *the subject* and *the duration*.

The DSA designs procedural obligations to deal with the restriction of content and establishes mandatory and clearer responsibilities for very large platforms. It addresses both canonical forms of content moderation and more subtle ways of making content inaccessible (for a detailed analysis on

shadow banning, see). In addition, the DSA prohibits platforms from organising or operating their online interfaces in a manipulative, deceptive, and distorted way. Similarly, platforms using recommender systems should allow users to change the parameters that influence recommendations.
To start the discussion on the system of liability, when a hosting service provider decides to remove or disable access to information, it should provide information about the removal of content as well as information on redress possibilities. These redress mechanisms consist of two steps. The first is an internal complaint mechanism, which can be followed by an out-of-court dispute settlement. The internal complaint mechanism allows users to challenge decisions about content removal or suspension and termination of their accounts. The DSA stipulates that platforms should handle complaints in a timely, diligent, and objective manner. Importantly, these decisions should not be made only by automated means. An online platform shall reverse its decision referred to without undue delay in case the information is not illegal *and if it is not incompatible with its terms and conditions*. It should be noted that Article 14 on Terms and Conditions requires digital media platforms to consider "freedom of expression, freedom and pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter". However, such a provision does not offer any substantive guidance on how fundamental rights should be implemented,[11] thereby leaving decision power to private companies on how to interpret normative matters (cf. Appelman, Quintais & Fahy, 2022). Article 18 stipulates that the out-of-court dispute settlement should be carried out by a designated body with demonstrated expertise, and which should be impartial and independent from both the online platforms and the recipients of the service. The out-of-court dispute settlement is binding on platforms, and a complaint can seek remedy in court. If this system can ameliorate the terms and conditions and the internal policies of the platforms remain to be seen.

The DSA imposes obligations on platforms to design a user-friendly mechanism to flag illegal content online. This is an already established practice among major platforms, but the DSA institutionalises a new entity, the one of *trusted flaggers,* whose notification should take priority. The status of a trusted flagger is awarded to entities (and not individuals) that pursue collective interests diligently and objectively and show particular expertise and competence in tackling illegal content. Lastly, the DSA also creates new actors at both the domestic and European levels. The digital service coordinator can investigate and enforce rules against digital platforms, whereas the European Board for Digital Service Coordinators has investigatory powers, advises the digital service coordinator, and develops European standards and code of conduct (cf. van Cleynenbreugel, 2021).

4. Transparency measures under the DSA

The DSA establishes a new and complex bureaucratic apparatus within and outside the main platforms to control how they operate. Strong transparency measures have become essential to give credibility to

[11] A similar provision is contained in article if the DSA rightly requires digital media platforms to tackle harmful content, it also remarks that "'harmful' (yet not, or at least not necessarily, illegal) content should not be defined in the Digital Services Act". Again, the EU is leaving to platforms to decide what is harmful or not, which is, to some extent also paradoxical, provided that their business model is designed to capture users' attention and as such it might very well benefit from the spread of harmful content as discussed above.

the entire European regulatory approach and to entrust platforms to decide current social arrangements and fundamental rights. To some extent, it seems that the more social media platforms are perceived as controllable, the less problematic it is to transfer authority to them (Maroni, 2022 2021). Transparency measures have thus far sat well with the platforms' strategy to seek forms of legitimation and deployed to convey the idea that there are trustworthy to avoid regulation while distracting from aspects of substantial accountability (Monika Zalnieriute, 2021). The final text of the DSA enshrines a strong system of transparency, which is constituted by progressively cumulative solutions. These span publicity, access to datasets, and disclosure of the use algorithmic to procedural fairness. Regardless of the granularity of some of these measures, a complete picture of the inner functioning of the platform is institutionally protected and hard to achieve: there is the risk of reproducing asymmetries of knowledge and foregrounding the influence of platforms over culture, rights, and democratic arrangements. To elaborate on these points, this section sketches and discusses some examples of transparency measures and couples them with critical transparency studies and critical data set studies.

Critical transparency studies point out how transparency is inherently connected with forms of neoliberalism (Adams, 2018); rather than providing clarity, transparency is a mechanism to manage visibilities (Flyverbom, 2016). Within this scholarship, transparency discourses are understood as a way of legitimising regimes of power while excluding other forms of governance (Adams Rachel, 2020), if not trading truth altogether (Koivisto, 2022). Critical dataset studies illuminate data as a socio-technical phenomenon and challenge assumptions that data can offer objective and accurate forms of knowledge (Boyd & Crawford, 2012). Given that transparency is a key element within the new EU regulatory regime, there might be merits in problematising narrow conceptualisations that tie transparency to control and *trustworthy* platforms. To be sure, verifying how platforms curate content and impact public discourse is a mounting democratic problem, and it is a given that having limited information sounds better than working on assumptions. The paragraphs below highlight that despite the DSA's transparency measures, there are aspects in platform governance and infrastructure that are not given to know. Furthermore, the quantity and diversified information needed to exert oversight should indicate that the solution to the social and individual harm caused by platform governance lies elsewhere—that is, in the business model, not in transparency solutions.

Although the DSA treats transparency problems as if they were numerically and qualitatively controllable, this work calls for caution in relying on transparency as an organisational pillar for content moderation, as the quality and accuracy of the information depend on the willingness of a given company to collaborate, as well as other mediating factors. For example, (pre-Musk) Twitter has been considered more advanced than other companies in making its data available to researchers.[12] On the contrary, Facebook has instead foreclosed access to data for researchers on different occasions,[13] provided flawed data thereby affecting the finding on disinformation.[14] Platforms themselves began

[12] 2020 Ranking Digital Rights Corporate accountability index (no date) Ranking Digital Rights. Available at: https://rankingdigitalrights.org/index2020/companies/Twitter (Accessed: October 30, 2022).
[13] DeLong LA, "Facebook Disables AD Observatory; Academicians and Journalists Fire Back - NYU Center for Cyber Security %" (NYU Center for Cyber SecurityAugust 21, 2021) <https://cyber.nyu.edu/2021/08/21/facebook-disables-ad-observatory-academicians-and-journalists-fire-back/> accessed October 30, 2022
[14] Papadopoulos L, "Facebook Fed Researchers Flawed Data, Caught by Its Own Transparency Report" (*Facebook Fed Researchers Flawed Data, Caught by Own Transparency Report*September 11, 2021)

enacting transparency measures in response to the criticism raised by civil society. The publication of their enforcement reports is currently a regularised practice; reports contain information about content removed due copyright's infringements, governmental requests, and requests to remove content based on local law privacy law (EU law or other national law). Although reports appear in different formats and labels in content removal vary across a range of platforms, the typology of information they publish is standardised, and it mirrors the way both infringements and threats have been conceptualised over time.

Far from being neutral, transparency reports are a quantification and categorisation mechanism that embodies early cyberlibertarian narratives. Transparency reports visually enable platforms to depict themselves as actively promoting human rights (Gorwa & Ash, 2020), and governments and regulations are conveyed as the main threat to internet freedoms (for an overview of this narrative, cf. Rikke Frank Jørgensen, 2017). Transparency reports give more visibility to governmental or law-based content removal requests and less to community guidelines enforcement (the first is displayed under three different categories as a sort of visual reiteration, whereas there is only one category for community guidelines enforcement) (Daphne Keller & Paddy Leerssen, 2020). However, civil society and academia have pointed out the major limitations of the reporting system: these reports are based on platforms' own assessments, and it is difficult to analyse the merits of the underlying cases (Daphne Keller & Paddy Leerssen, 2020).

To overcome the problem of platform grading their own homework, transparency measures are progressively acquiring new dimensions: Transparency is put into practice as publicity, quantification, and categorisation, as well as procedural fairness. Over time, platforms have overcome their initial reluctancy about disclosing their internal rules, and today, they publish their community standards to provide a range of information about which content is allowed or not. Transparency in content moderation practices owns much of its appeal to the promise of holding platform power into account. However, the bulk of platform operations are carried out algorithmically, whose code and decision-making criteria (i.e. how they prioritise, classify, associate, and filter information) are not public (Diakopoulos, 2014). Platforms are private actors, meaning that they are legally protected to limit what information to disclose and this either because of trade secrets, or risks of gaming the system or to protect users' privacy (Diakopoulos, 2014).

In this regard, access to datasets has been considered critical to understanding how content travels and to studying how people encounter, share, and interact with online disinformation. Although a detailed discussion of datasets falls outside the scope of this paper, big-scale data are considered difficult to analyse; thus, access to datasets is more promising in different and small contexts than in big digital media platforms. Furthermore, Allen et al. 2021pointed out that the data can be limited or biased in a way that might lead to distorted conclusions. This happens more easily with big-scale data, in which noise is added to protect users' privacy, and where some data can be over-represented (e.g. entertainment) and other underrepresented (gaming), and this bias affects the relationship between viewing, clicking, and sharing. One example of this has been Facebook's under-representation of micro-targeted ads containing links to political fundraising (Allen et al., 2021). Similarly, researchers

<https://interestingengineering.com/culture/facebook-fed-researchers-flawed-data-caught-by-transparency-report> accessed October 30, 2022

investigating the AirBnB dataset concluded that the data released by the company were limited and that it is important to equip transparency initiatives with auditing (Cox & Slee, 2016).

As argued, transparency mechanisms are tied to the institutionalisation of platform private power. In this regard, *procedural fairness* is also encroached upon with the call for greater transparency. In content moderation debates, it has been underlined that out-court mechanisms, such as the Facebook Oversight Board, have been established with the view of opening the implementation of its community standards (Klonick, 2020) and as a justificatory mechanism (Kettemann & Fertmann, n.d.). However, doubts about the capacity of the Facebook Oversight Board to open Facebook decision-making processes remain and are acknowledged by the board itself on several occasions (*Oversight Board Demands More Transparency from Facebook*, n.d.). On Trump's decision, the board remarked that:

> "Unfortunately, the lack of transparency regarding these decision-making processes appears to contribute to perceptions that the company may be unduly influenced by political or commercial considerations" (Facebook Oversight Board, 2021; Marta Maroni, 2019).

The DSA codifies, harmonises, and tries to complement the scope of transparency standards outlined above. It also accommodates the need to balance the private nature of digital media platforms with aspects of public interest and to establish a 'tiered system of transparency' (McCarthy, 2020). This balancing exercise translates into making some information open while restricting access to others. Information about content removal is still widely accessible, even if it requires interpretation. On the contrary, when the information touches upon the business model of platforms, it is both restricted and institutionally mediated and presupposes specific and technical competences to be interpreted.

In many respects, transparency is primarily designed *as publicity*. Article 14 establishes that the Terms of Conditions should be written in a clear, unambiguous, and easily accessible format. Large companies are also required to provide graphic material to understand the terms, as these are unreadable for most human beings, given their length and technical language (Nicholas LePan, n.d.). However, as Suzor pointed out, the terms of service are take-it-or-leave-document (Suzor, 2019); therefore, individuals are powerless in negotiating the use of the services, and it is left to regulatory intervention to correct the content of terms of services. Further, Article 14 mandates providers to make public the policies, procedures, measures, and tools used for the purpose of content moderation, including algorithmic decision-making and human oversight.

To continue on transparency as publicity, the DSA imposes on providers of intermediary services the obligation to publish comprehensible and detailed reports on their moderation practices (Article 15). The degree of transparency varies according to the size of the platform. These reports should make available quantified and categorised information about content removal practices, such as the number of orders received from Member States and the number of notices received through the notice and take down mechanism. It should be noted that the DSA puts emphasis on content moderation measures initiated by the provider itself. This would need to disclose meaningful and comprehensible information about:

> "the use of automated tools, the *measures taken to provide training and assistance to persons* in charge of content moderation, 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 through the service, and other related restrictions of the service; the information reported shall be *categorised by the type of illegal content* or violation of the terms and conditions of the service provider, by the detection method and applied by the type of restriction applied". (*Italics are mine*)

According to Article 24, digital platforms have extra obligations to report the number of disputes submitted to out-court-dispute settlement, the number of suspensions, and the number of suspensions imposed against misuse (i.e. suspension of the service to users who frequently provide manifestly illegal content). In addition to the abovementioned obligations, very large online platforms (Article 42) are required to provide information about the human resources employed, their qualifications, and their linguistic expertise. This provision addresses aspects of the *enforcement* problem in the policy debate, for example, questions of who the moderators are, whether there are enough of them, and whether they are apt with the task of evaluating content removal. Finally, platforms are expected to publish their reports every six months and transmit them to the digital services coordinator for the audit implementing reports. Reports can omit information that can cause significant vulnerabilities to the security of platform services, undermine public security, or may harm recipients. However, complete reports containing the justifications for the removed information should be submitted to the digital services coordinator (Article 33). Under the DSA, individuals are also entitled to transparency (*personal transparency*). This type of transparency enables users' agency in interacting with ads and when their content is restricted or removed. 'Personal transparency' will eventually become public when it is collected, aggregated, and transformed into accessible information for public scrutiny.

Platform users are not often aware that they are interacting with sponsored content, and Article 26 stipulates that users should be able to recognise an advertisement, receive information about the sponsor, and access the main parameters used for targeting. According to Article 27, platforms should disclose the parameters used when they use a recommender system; they should also grant users the possibility of modifying them. Further, providers of very large platforms using 'recommender systems should provide at least one option for each of their recommender systems which is not based on profiling' (Article 38).

The DSA generally prohibits displaying advertainments based on special categories of data, such as data revealing racial or ethnic origin, health, and sexual orientation (for more cf. art 9 of the GDPR). Very large platforms have the additional obligation to make publicly available through the application programming interface a repository containing very specific information about an *advertisement*, including where the ad targets a specific group. The DSA considers options to empower users to interact with recommender systems. However, this is an individual-centred solution, and the information disclosed is restricted to ads only, and it is unclear how it can address structural problems such algorithmic discrimination, inequalities, and the questions of media pluralism. Furthermore, one might wonder the extent to which modifying one criterion when interacting with a recommender system leads to substantial changes in the way recommender systems operate.

Transparency intersects with *procedural fairness*. The users whose content has been removed should receive a *reasoned statement*, which contains contextual information, legal, and contractual grounds, and information about redress mechanisms (Article 15). The DSA also addresses the problem of making publicly accessible the justifications subsuming a decision on content removal, as legal arguments are critical for assessing the reasoning of decisions and the legitimacy thereof. The DSA requires the EU Commission to maintain a database to gather the decisions and statements of reasons for removing and restricting content (Recital 66). This would allow scrutiny of patterns in the decision-making system about content. However, content moderation happens on a high level of rate and scale, and the volume of information that the database will collect and generate raises questions about the real possibility of qualitatively analysing the decisions. There is no algorithmic transparency, that is, the obligation to reveal the inner logic of the algorithmic in the DSA, because of the need to protect the way algorithms function. Rather, as discussed in the previous sections, there is an obligation to

disclose when algorithmics are deployed. In this context, the closest source to study patterns in consumption and dissemination of content then becomes access to datasets.

Article 40 requires platforms to make available access to their datasets. Their use is restricted to researchers and the Commission, which can access the data for "sole purpose of conducting research that contributes to the identification and understanding of systemic risks" and to assess "the adequacy, efficiency and impacts of the risk mitigation measures". It should be noted that the EU Commission (and not researchers) could also request "very large online platforms to explain the design, the logic, the functioning and the testing of their algorithmic systems, including their recommender systems". This additional information provides a means for better analysing the vast amount of information contained in the datasets.

If policy debates have pointed out that unconditional access to datasets comes with privacy concerns and the risks of gaming the system, other means could have been envisaged to limit the possibility of harm (e.g. a system of strong authentication for users or quota to prevent bots from exploiting the platforms and its content), rather than restricting access, thus impairing the possibility to scrutinise the dataset from multiple and analytical points of view. If, however, the data are central to gathering "massive volumes of data about given platform's users and content" and to studying information flow (Tromble, 2021), as also discussed above, the data would not be neutral; they would be mediated and formatted by the platforms, which could aggregate data to make the analysis more difficult, for example, by providing information that is not needed. In addition to mediation inherently inbuilt into the dataset, there is also the role of the digital service coordinator (i.e. a domestic authority) to consider, as this is the one acting as a sort of trustee and approving the status of vetted researchers based on their affiliation, conflict of interests (independent from commercial interests), and their research proposals. The latter should demonstrate "that their access to the data and the time frame requested are necessary for, and proportionate to, the purposes of their research, and that the expected results of that research".

This provision appears problematic in terms of academic freedom, as the digital service coordinator will act as gatekeeper for the "vetted researchers", and can also terminate access to the data based on 'following an investigation either on its own initiative or *on the basis of information received from third parties'*, in case the researcher no longer meets the conditions "to be vetted". This provision could be activated to interfere with ongoing research. Finally, against all the mediation, control, and restrictions discussed above, it is also worth considering that while conducting their analysis, researchers are themselves driven by different objectives and consequently clean their data by picking different variables that necessarily influence the outcome of the research.

Overall, the construction of this article exhibits several compromises made at the EU level in balancing the freedom of private companies and access to information that is important for the public interest. These compromises create privileged forms of access to knowledge in a field that already requires very specific competences. Thus, it is also worth noting that platforms can refuse to share their data if 'giving access to the data will lead to significant vulnerabilities in the security of their service or the protection of confidential information, in particular trade secrets' (Article 40). This phrasing mirrors what Tromble defined as the "property black-box problem"—that is, "the platforms are under no obligations to reveal whether, why, and how some data are made available via an API while other data are not" (Tromble, 2021).

At the moment of writing, the delegated acts are not available; thus, in the future, it would be important to analyse how different rights are going to be reconciled. In particular, *trade secrets, security of platform service, and privacy* should not be utilised in a way that restricts access to information important for public scrutiny and debate, as these rights do not necessarily play one against each other and can be simultaneously implemented. The DSA also envisages an auditing system and establishes

substantial sanctions, up to 6% of the annual income or turnover, in case intermediary services fail to comply with their obligations, and creating extra penalties if an intermediary supplies incorrect, incomplete, or misleading information or fails to reply or rectify incorrect, incomplete, or misleading information.

5. Conclusions