SciencesPo
SCHOOL OF PUBLIC AFFAIRS

PUBLIC POLICY MASTER THESIS

April 2025

Assessing the Digital Services Act's effectiveness in fostering meaningful transparency in social media platforms

The case of Meta

Lola Pottier

Master's Thesis supervised by Caterina Froio

Second member of the Jury: Sylvain Parasie

Master in European Affairs
Policy Stream Europe in the World

Abstract

Against the opacity and asymmetrical power that characterise digital platforms, the European Union introduced in late 2020 the Digital Services Act (DSA), an ambitious legislation that seeks to promote transparency and accountability in platforms. While the directive's potential benefits and pitfalls have been widely discussed, few studies evaluate how it has translated in practice. This study is a first step towards bridging this gap. Taking Meta as a case study, it assesses to what extent the DSA has been effective in fostering meaningful transparency in social media platforms. It compares Meta's past practices with its current implementation of the DSA, investigating documents produced by Meta, 625M statements of reasons from the Transparency Database and Instagram's user-facing features. The results indicate that the DSA has increased meaningful transparency at Meta, but not equally across individual, aggregate and systemic levels, the three definitional layers of meaningful transparency. While procedural user-facing transparency is quite robust, the DSA has more difficulty fostering understanding of the platform's overall content moderation patterns and impact on society. This can be explained by shortcomings in the DSA itself, but also by the visibility-management strategies that Meta deploys when releasing public information, which hinder its clarity. The thesis concludes by issuing seven recommendations covering two areas: raising the minimum of what is expected of platforms and promoting independent research.

Key words

Transparency, Content Moderation, Digital Services Act (DSA), Meta, Social Media, European Union

Acknowledgements

As I am wrapping up these intense three months, it is important for me to thank all the people who helped me successfully complete this work.

First of all, I would like to thank my supervisor Professor Caterina Froio for consistently providing me with helpful, considerate, and rigorous advice when I needed it.

I would also like to thank the second member of my jury, Professor Sylvain Parasie, for accepting to spend time reading dozens of times the word “meaningful” and discussing the DSA with me.

High in my thoughts are my friends and family, who have provided me with constant support throughout the thesis. Whether you contributed to technical coding discussions, passionate debates about social media, or simply gifted me with a sympathetic ear, know that you were instrumental in the accomplishment of this work.

List of abbreviations

AI - Artificial Intelligence
DMA - Digital Markets Act
DSA - Digital Services Act
EU - European Union
FOI - Freedom of Information
SoR - Statements of Reasons
TDB - Transparency Database
ToS - Terms of Service
TR - Transparency Report(s)
VLOP - Very Large Online Platform
VLOSE - Very Large Online Search Engine
US - United States (of America)

List of figures

Figure 1: The four definitions of transparency…………………………………………… 10
Figure 2: Summary of what meaningful transparency means in content moderation…… 14
Figure 3: Types of intermediary service providers under the DSA……………………… 16
Figure 4: (main) Transparency provisions expected from intermediary service providers under the DSA…………………………………………………………… 17
Figure 5: Path of a content takedown notice on Instagram……………………………… 25
Figure 6: Daily volume of moderation actions across Meta platforms …………………. 30
Figure 7: Automated detection rates across Meta platforms…………………..……… 31
Figure 8: Daily moderation actions on Facebook and Instagram split by content type…. 33
Figure 9: Daily moderation actions on Facebook split by category……………………… 35
Figure 10: Comparison of aggregate Article 16 notices across the transparency report (TR) and transparency database (TDB) for Facebook and Instagram between 01/04/2024 and 30/09/2024………………………………………………………… 38
Figure 11: Comparison of selected moderation actions across the transparency report (TR) and transparency database (TDB) for Facebook and Instagram between 01/04/2024 and 30/09/2024………………………………………………………… 39

Table of content

Main contribution.. 1
Introduction..2
1. Interdisciplinary state of knowledge.. 3
1.1. The platform: a particular business model and a purported neutrality...3
1.2. The trouble with content moderation on social media platforms..5
1.2.1. A necessary component of platforms...5
1.2.2. A great power that comes with great opacity..6
1.3. Is transparency the solution? Interdisciplinary perspectives...7
2. Making transparency transparent: definitions and theoretical framework................................9
2.1. Defining transparency: a four-layer framework...9
2.2. Defining meaningful transparency in the context of content moderation on social media platforms......11
3. The DSA: introducing transparency to mitigate the undue power of platforms...................... 14
3.1. Background: the need for a modernised framework to regulate digital services.................................... 15
3.2. A tiered regulation centered around rights and transparency...16
4. Meta's historical approach to content moderation and transparency...................................... 18
5. Methodology, data and sources..20
6. Understanding moderation sanctions at the individual level..24
6.1. Procedural rights: explaining "whether" and "why"...24
6.2. Enforcement procedures: explaining "how"... 26
7. Understanding moderation sanctions at the aggregate level...28
7.1. The Transparency Database: a unique but incomplete overview of sanctions..29
7.1.1. Exploring the Database: four lessons about Meta's moderation...29
7.1.2. A flawed and obscure glimpse of reality.. 32
7.2. Transparency reports: a case of obscure transparency... 36
7.3. Comparing the database and the reports: a boost in meaningful transparency....................................... 38
8. Understanding moderation at the systemic level..40
8.1. Towards advertising and recommender systems transparency...40
8.2. Exposing social media's structural impacts: innovative but still limited oversight................................ 43
Conclusion and recommendations...45
Bibliography.. 50
Annex I - Overview of the Transparency Database.. 73

Main contribution

The Digital Services Act (DSA) is an innovative piece of legislation, several of whose provisions are unprecedented. One year and a half after it first entered into force for Very Large Online Platforms and Search Engines, few studies have yet studied what its effective impact has been on fostering transparency in social media platforms. Hence, contrary to most of the existing literature on the DSA and transparency, this thesis goes past theory and studies the DSA in practice, considering the case of Meta.

While the literature on social media transparency tends to focus on individual features, like recommender systems or advertising transparency specifically, this work instead seeks to provide a comprehensive overview of Meta's practices on transparency. It reviews the main provisions of the DSA point by point and compares their current implementation with past practices. By doing so, it offers both a thorough understanding of the different articles of the DSA and how they translate in practice, and a historiographical account of Meta's transparency practices to this day.

To get to this result, this study further makes two contributions through its methodology. First, it puts forward a framework that defines and operationalises meaningful transparency in the context of content moderation on social media, by combining the existing disparate inputs of the literature into one coherent whole. The goal of this definitional work is to contribute to the literature by conceptualising not so much the meaning of transparency, which is by now quite extensively discussed, but of *meaningful transparency*, whose theorisation is scarcer. This provides a basis for analysing social media transparency. Second, it uses data that is not yet studied by academia. In particular, to the author's knowledge, no study has yet been published that investigates the Transparency Database after August 2024, and after January 2024 for Meta specifically, nor is there published work that studies the 2024 transparency reports and compares the two provisions. Trujillo et al. (2024) had undergone such a project in the first months of the database, but the period that they studied did not match with the report's reporting period so their comparison could not be complete. Further, this thesis studies the database over eleven months and includes non-VLOPs, which has not been done so far. In addition to being useful for the analysis, this contribution provides a glimpse of the database for audiences who are not technically able to investigate it, and sets in stone the functioning and flaws of the database before the upcoming Implementing Act changes it.

Additionally, the thesis points to specific elements of Meta's implementation of the DSA that are surprising. This could be of interest to regulators, researchers and civil society, to better understand Meta and improve the DSA for the next regulatory periods.

Finally, it is worth mentioning that at the time of writing, profound structural changes seem to be happening globally, including in the digital sphere. Amidst transatlantic tensions, the DSA has taken an awkward position, being regarded by some as a potential EU weapon against US tech giants (offensively or defensively). This thesis is an opportunity to scientifically assess and establish the facts that stem from the DSA's implementation, to shed light on what the DSA does and does not do. This way, it can inform public debate with rigorous evidence.

Introduction

Whether we like it or not, social media has become central to social exchanges. The biggest social media platforms are now billion-dollar corporations hosting billions of users, who exchange a gigantic amount of posts, comments and likes every day. At the same time, social media companies have been so far governing their product with relative discretion. We know, yet we don't always realise, that social media have built algorithms and systems that select, prioritise and order content for users, that they make revenue from the visualisation of particular types of content as part of their advertisement-based business models, and that they have drafted moderation rules and built mechanisms to enforce them more or less consistently. Yet we don't know, for the most part, how they proceed, why and for what purposes. In this situation, isolated users (by design), but also the societies in which they are embedded, find themselves powerless against large profit-led multinationals whose systems select and curate the information that they create and consume while acting as potential massive amplifiers of specific pieces of content. This grants social media platforms significant power, that they have so far exercised in a particularly opaque and unaccountable manner. This can prove to be a serious problem for democratic fora.

Against this backdrop, the European Union (EU) introduced the Digital Services Act (DSA) in late 2020. The directive represents an unprecedented attempt, by its scale and ambition, to promote transparency and accountability in platforms on their content moderation and risk management. It was presented as a way to "open the black box" of social media algorithms and to put an end to the "so-called 'Wild West' dominating [Europe's] information space" (Breton, 2022). This emphasis on transparency, which permeates Western thinking, has been praised (who would be against transparency?), yet this simple-minded understanding has also been criticised: if a system is so hard to understand that it is comparable to a black box, making its sides transparent will only reveal a pile of intricate and unintelligible systems (Rieder and Hofmann, 2020). In this context, some have introduced the concept of meaningful transparency to refer to the ability to understand the information that is being communicated.

Nearly one year and a half after it first entered into force in 2023, there are legitimate questions as to what the actual impact of the DSA has been on making social media platforms not only more transparent, but more *meaningfully* transparent in the sense that it has fostered additional understanding of how they operate. This work seeks to be a first step in undertaking such research. Taking Meta as a case study, this thesis asks the following research question: **How effective is the DSA in increasing meaningful transparency in social media platforms?**

Comparing past practices with current implementation, we find that the DSA has positively impacted the degree of meaningful transparency that can be observed at Meta across individual, aggregate and societal levels, without yet managing to shed light equally on all processes. We demonstrate that it guarantees robust meaningful transparency at the procedural user-facing level, though processual explanation can lack, but it has more difficulty fostering meaningful transparency on the aggregate moderation patterns of platforms, as well as on the systemic impact of their moderation on society. We identify two sources to such impediments. First, blind spots and inherent limitations of the DSA itself can in practice lead to lower-than-expected results. Second, we observe a tendency for Meta to engage in visibility-management strategies when the information that is released is public, as opposed to directed to a single user, which leads it to comply to the minimum extent possible and/or to produce material whose informational value is limited. Of course, these observations take stock of the situation as of April 2025, but they could change as the DSA continues to be implemented.

This analysis is structured as follows. Section 1 presents the current state of knowledge on the key topics of this thesis with an interdisciplinary perspective, before Section 2 focuses on defining and operationalising the notion of "meaningful transparency" in the context of content moderation on social media. Section 3 and Section 4 present background information on the DSA and the chosen case study, Meta, in relation to transparency. Section 5 then presents the data and methodology used to conduct the study. Finally, Sections 6, 7 and 8 each focus on a definitional layer of meaningful transparency and analyse the relevant DSA provisions in relation to their implementation by Meta and their corresponding past practices. The final section concludes and puts forward seven recommendations to further increase, and guarantee, meaningful transparency in social media platforms.

1. Interdisciplinary state of knowledge

This section surveys the academic literature to establish the context behind the DSA and its emphasis on transparency. The review is organised around the three core themes of this thesis: the platform, content moderation, and transparency.

1.1. The platform: a particular business model and a purported neutrality

Originally, the word "platform" refers to the technological infrastructure that supports the creation of digital applications (Gillespie, 2010). This approach permeates computer science, where attention is given to the hardware on which platforms operate, but also to their (re-)programmable software systems (Poell et al., 2019). A similar attention to architecture can be found in business management: Baldwin and Woodard (2009) characterise platforms as systems interweaving stable components with evolvability. In economics, more importance is given to the function of platforms in the market than to their inherent characteristics. They are regarded as intermediaries between groups of providers and consumers, i.e. "digital infrastructures that enable two or more groups to interact" (Srnicek, 2016, p.24). They are a new organisational form that structurally changes economic exchange by flattening hierarchies into networked interactions (Acs et al., 2021), and is believed to increase trade efficiency by significantly reducing the cost of information sharing and matching economic actors better and faster (Xue et al, 2020). More critical definitions across social science disciplines emphasise the inherent capitalist rationale of platforms and their emphasis on commodification and data collection (Zuboff, 2019; McMillan, 2020; Sander, 2020). While all relevant, these definitions do not alone encapsulate the complex model that characterises today's platforms. Accordingly, Gillespie's (2018) definition, which combines those different disciplinary perspectives, is the one that will be preferred in this work. To him, platforms are:

> Online sites and services that [1] host, organize, and circulate users' shared content or social interactions for them, [2] without having produced the bulk of that content themselves, [3] built on an infrastructure for processing data for a range of different purposes including the generation of profit, and [4] which moderate the content and activity of users (Gillespie, 2018, p.18-21).

This definition points to three features of platforms that the reader should be aware of. First, since platforms are recognised as "hosts" of content that they have not produced, they convey an image of *neutrality* (Chander and Krishnamurthy, 2018; Roberts, 2018). In fact, platforms have specifically labeled themselves this way to be seen as neutral (Gillespie, 2010), even if they are more varied than this sole label suggests (Caplan, 2018). The positioning is strategic: by doing so, platforms not only appeal to users through an attractive promise but they also deny any responsibility from the content that they host, including from a legal perspective. Indeed in US (and to some extent EU) law, publishers bear the responsibility of their own content but internet intermediaries do not, for the most part (Tyler et al., 2025). As rather new markets, platforms have also been loaded with the liberal and neoliberal imaginary of a neutral and apolitical market (McKee, 2017). Yet platforms are not neutral. Platforms "intervene" (Gillespie, 2015): they guide, distort or facilitate social activity (Gillespie, 2015) through their design (Duquenoy, 2005), technological infrastructure (Leerssen, 2020), policies (Suzor, 2018) and practices (Roberts, 2018). They are driven by specific values (Hallinan et al., 2022), which are influenced by the particular cultural setting from which they originate (Davis and Xiao, 2021). Hence, when analysing platforms, it should be remembered that platforms have agency.

The second feature constitutive of platforms is the particular economy and business model(s) in which they are embedded. Most platforms, especially social media platforms, work on an advertisement-based business model (Acemoglu et al., 2024), whereby they supply two products to two customer groups: "content to readers (in exchange for readers' attention) and readers' attention to advertisers (in exchange for monetary payments)" (Newman, 2019, p.1527). In this attention economy, human attention is considered a scarce commodity (Crogan and Kinsley, 2012) that platforms, as attention brokers, resell (Wu, 2019). Yet attention is not the only currency to be traded in this way. Data, central to Gillespie's definition, is at the core of platforms' activities and often constitutes another currency (McMillan, 2020) for users to pay what can be considered a "rent" for being on the network (Langley and Leyshon, 2017). In practice, platforms extract value out of the quantification of human life, a process known as *datafication* (Mejias and Couldry, 2019). Therefore, while platforms are commonly thought of as offering "free" products, they rather offer zero-price products whose costs are not necessarily inferior to those of other markets (Newman, 2019). This can be obfuscated by the current *platformisation* of society, the process by which platforms are gradually instilling all spheres of life, recalibrating common imaginaries around their model (Poell et al., 2019). Hence, when analysing digital platforms including for apparently unrelated elements such as transparency, it must not be forgotten that these products are designed to serve specific goals, which are often, like Gillespie's definition suggests, profit seeking.

The third element to consider is content moderation, which will be discussed in the next sub-section. While a variety of platforms exist, the focus is put from now on on a particular type of platforms, social media, which are defined as "Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others." (Carr and Hayes, 2015, p.50).

1.2. The trouble with content moderation on social media platforms

The most common conception of online content moderation is that of *sanctioning*, referring to "mechanisms that aim to prevent harm by removing or reducing the visibility of rule-breaking content" (Drolsbach and Pöllochs, 2023, p.1). Though it is a tempting one, this restriction obscures the other ways in which platforms act on content. Not only do platforms sanction, but they also *order*: they are both content gatekeepers and content organisers (Sander, 2020). Therefore, content moderation should be defined as "the process in which platforms shape information exchange and user activity through deciding and filtering what is appropriate according to policies, legal requirements and cultural norms" (Zeng and Kaye, 2022, p.80). Moderating content on social media is arguably a daunting task. It deals with a high volume of content and is characterised by a plethora of tensions, starting with the fact that it is both a defining feature of platforms and an extremely complex mechanism that grants platforms significant asymmetric power over users. Exploring this tension allows us to understand why content moderation is an intricate practice and why it has been attracting the interest of researchers and regulators alike in the last years.

1.2.1. A necessary component of platforms

One of the reasons why content moderation has become a central question for academics, legislators and companies, is that it is fundamentally indispensable on social media platforms. On a practical level, content moderation is essential to guarantee the system's usability. First, platforms are constrained by limited material and architectural capacity (Langvardt, 2018), for example fixed-sized servers or data centers. Though it is especially true of early internet times (Zuckermann and Rajendra-Nicolucci, 2023) or today's nascent platforms, moderating content is a way to avoid the system's overload (Grimmelmann, 2015). Second, some degree of sorting and hierarchisation is necessary to avoid *cacophony*, a situation where it would be impossible to absorb information because everyone would be talking at the same time (Grimmelmann, 2015). Not only would this make social media pointless, but it would also disincentivise users from spending time on the platform, which is yet the basis of current social media's business model. Likewise, without some moderation, platforms would likely be filled with spam and graphic content, which would make the service unusable.

On a more substantive and perhaps normative level, content moderation can be considered necessary to prevent illegal and/or harmful content from getting promoted and shared. This includes content ranging from terrorist content and child pornography to hate speech and cyberbullying, all of which social media platforms have been found to host (Tyler et al., 2025). While moderating illegal content can be accepted, there is less consensus on moderating “lawful but awful” content, which some, starting with Meta’s CEO Mark Zuckerberg (Zuckerberg, 2025), have denounced as censorship (Citron and Penney, 2024). Assessing what online harm means is especially difficult, as it depends on varying societal moral norms, markets and architecture (Nourooz-Pour, 2024). For example, the United States tends to have a more individualised approach to free speech, where the state is expected to never interfere, while Europe is generally more inclined to recognising legal limits on speech (Douglas-Scott, 1999). However, what is often seen as dependent on a society’s values may actually be a universal human value. In the Universal Declaration of Human Rights, Article 19 grants individuals the right to freedom of opinion and expression without interference, but Article 29 also recognizes that this right can be limited to secure respect for the rights and freedoms of others (Mansell et al., 2025). If we consider the case of hate speech, which refers to any offensive material deliberately degrading a particular group (Ștefăniță and Buf, 2021), empirical evidence showed that it has tangible psychological consequences on victims, while also deterring them from exercising their freedom of speech (Citron and Penney, 2024; Kosters and Gstrein, 2024; Langvardt, 2018). Further, it has been observed that online hate speech and offline violence are correlated (Relia et al., 2019; Siegel, 2020; Wilson and Land, 2021). Considering that hate speech disproportionately targets the same groups (Castaño-Pulgarín et al., 2021), for example the LGBT community (Ștefăniță and Buf, 2021; Walters et al., 2020), moderating hate speech (and similar “awful” content) can be considered a way to protect freedom of expression (Citron and Penney, 2024) on society level, while potentially protecting the integrity of individuals.

1.2.2. A great power that comes with great opacity

Yet whatever restricts speech or sorts information inherently has an influence over the rights to freedom of expression and pluralism of information. This is why rule-setters and enforcers of moderation actions necessarily have power over the moderated. Still, how this power is handled depends on the moderator’s identity (values, philosophies) and technical choices (style, actions) (Jiang et al., 2023). Depending on how the moderator solves the dilemmas of content moderation, e.g. efficiency versus quality, centralisation versus distribution (Jiang et al., 2023), it is possible for this power to be exercised in a balanced way.

Specifically, most US social media platforms benefit from a blatant asymmetry towards users, not only because they strip users off their personal data or impose their terms of service (ToS) (Poell et al., 2019), but also because they alone control and understand their moderation process. This means that the inherent power that comes with moderating content is exercised in an asymmetric way, which, scholars denounce, has very much to do with the inherent opacity of social media platforms. To begin with, the rules that platforms use to act on content have been largely opaque. It took a long time for social media platforms to acknowledge that they were moderating content, even though they have been since their creation (Citron, 2018), and an even longer one for them to publish their ToS. Yet ToS do not account for the implicit rules (Sander, 2020), nor do they include how they have been created (Tyler et al., 2025; Gorwa et al., 2020), why they change, and how they are applied (Kosters and Gstrein, 2024). Further, ToS have been observed to be unequally enforced. People with public influence or financial power have been able to get favourable treatment (Sander, 2020), while marginalised communities like native American people (Vaccaro et al., 2021) have faced harsher treatment. Not all platforms are the same, though: in general, large platforms are more opaque than smaller ones (Urman and Makhortykh, 2023), and commercial-moderated platforms more opaque than user-moderated platforms (Cook et al., 2021). In any case, these shortcomings might increase in frequency and severity given the surge in the use of artificial intelligence (AI) in moderation. Because the rarity of some cases means that less training data is available, AI algorithms could make mistakes that would induce algorithmic discrimination (Palmeira Ferraz et al., 2024) and raise justice issues, while increasing opacity (Gorwa et al., 2020). Scholars also express concern over overreliance on AI (Kosters and Gstrein, 2024), which would depoliticise content moderation (Gorwa et al., 2020) and further strip meaning off language by ignoring context (Wilson and Land, 2021). All these elements raise accountability (Nourooz-Pour, 2024; Rieder and Hofmann, 2020) and legitimacy (Palmeira Ferraz et al., 2024) issues, which are exacerbated by the fact that the "monopoly" of moderation actions is held by private companies (Palmeira Ferraz et al., 2024, p.381). In fact, Roberts (2018) argues that content moderation is specifically designed to increase platforms' profit. This is consistent with the tension identified in the previous part between the neutral platform and the deeply asymmetric platform seeking to maximise profit. Therefore, there are good reasons for undertaking some action to correct this undue power, as the status quo could pose serious democratic problems. As Olesen (2025) rightly points out, though, this is not to say that other organisations are not opaque but only that "Big Tech companies are opaque and socially and democratically consequential in new, rather *specific* ways that (...) are distinct enough to merit discussion in their own right" (p.12).

1.3. Is transparency the solution? Interdisciplinary perspectives

Given the imbalance that currently characterises yet essential content moderation on most social media platforms, it seems pressing to identify solutions to restrain their power and correct this shortcoming. To this end, several suggestions have been made. To name a few, Sander (2020) proposes a human-rights based approach to content moderation that draws on the UNGP framework, while Suzor (2018) and others propose *digital constitutionalism*, a framework that seeks to limit governance power by imposing processual obligations that place legitimacy and accountability in the center. A common feature that these proposals share, and which has been a core element of platform governance over the last years, is a focus on *transparency* as a remedy against the opacity of platforms. This fondness for transparency is not a twentieth-century invention: Enlightenment thinkers were fascinated with the concept, and 17th-century philosophers praised a very similar value that they called "sincerity" (Ward, 2017). This endeavour is not unique to platforms either, thus it is useful to zoom out and look at how transparency has permeated other disciplines.

Transparency was widely popular in the 1980s-2000s, with the worldwide advent of Freedom of Information (FOI) laws (Meijer, 2014), which mandate governments to release documents upon request. They were seen as granting citizens a new right to information by shedding light on an opaque system. Over the years, transparency principles have also been applied in the regulation of other sectors, including finance, environmental regulation and political campaigning, so that transparency is extensively discussed across domains in the literature. Ackerman and Sandoval-Ballesteros (2006) identify a range of benefits fostered by transparency across three areas (politics, economics and public administration), in the context of open government. On the political level, they believe that transparency could make citizens more aware and involved in the activities of their representatives, and could also respond to an accountability deficit. A similar argument is advanced for social media platforms: it highlights the importance for users to understand the governance in which they are embedded (Gorwa et al., 2020), which impacts their ability to foster change within platforms (Urman and Makhortykh, 2023) and to sanction them, e.g. for not being network-neutral enough (Pappas et al., 2015). On the economic level, transparency increases efficiency by allowing more informed decisions, that is, by reducing information asymmetries (Matheus and Janssen, 2019). Finally, on the administrative level, transparency could lower corruption and increase trust in government. To some extent, the latter argument is validated in the corporate social responsibility literature, since transparent communication is found to increase consumer trust in the company (Kim and Lee, 2018; Lee and Chung, 2023). Positive impacts against corruption were also observed in the context of political campaigning (Wood, 2021). In the context of platforms, transparency is believed to be able to foster trust in the platform (Suzor et al., 2019), which the empirical experiment led on e-commerce platforms by Veltri et al. (2023) suggests is true. Finally, two major arguments in favour of transparency for social media platforms, which are necessarily absent from discussions on FOI laws, are first the importance for researchers (Suzor et al., 2019) to monitor the potential detrimental effects of content moderation (Sander, 2020), and second to acquire knowledge on the functioning of platforms to better inform legislators and draft appropriate regulation (Leone de Castris, 2024).

In recent years, critical transparency studies have made increasingly salient that transparency only represents the simple solution that society craves for to solve its ills (Koivisto, 2019), yet transparency is not simple and is not, alone, the solution to all ills. To start with, transparency's effectiveness depends on context (Uras, 2020) and should not be considered a silver bullet. In addition, transparency is filled with assumptions that are not necessarily correct, which makes the aforementioned benefits ill-advised. A first example is that transparency posits that actors will respond rationally to the information that is disclosed, but it is not necessarily the case (Annany and Crawford, 2018), as the case of climate finance shows (Ameli et al., 2019). Another example shows how transparency is limited and can be harmful (Annany and Crawford, 2018): Obama’s presidency has demonstrated that too many promises of transparency creates too high expectations that will run up against the impossibility of demonstrating full transparency (Coglianese, 2009). Applied to companies, including platforms, transparency is found to be a form of visibility management (Wagner et al., 2020) that can be exploited for marketing and branding purposes in what is essentially transparency washing (Zalnieriute, 2021; Padfield, 2025).

However, researchers should be careful not to fall in ethics bashing when they criticise ethics washing (Bietti, 2020). Despite its faults, transparency may constitute part of the solution to address the opacity of social media platforms, such as a first step towards more robust legislation (MacCarthy, 2022). It at least deserves an inquiry that considers it seriously. Furthermore, discussions around transparency often forget to *define* what transparency is, so that it is unclear whether various praises and criticisms are dealing with the same object of analysis. The next section deals with this issue.

2. Making transparency transparent: definitions and theoretical framework

While transparency is extensively mobilised in various discourses (Schlag, 2023), the word itself is seldom defined (Fox, 2007). This makes it a convenient all-purpose term on which everyone agrees despite projecting different understandings (Fox, 2007), which is a problem for any analytical work on the subject. At the same time, Forssbæck and Oxelheim (2014) underline that the broadness of the term makes a "strict and universally viable definition virtually impossible" (p.5). Without taking on such comprehensive intentions, this section attempts to summarise theoretical contributions on the definitions of transparency, before discussing their application in the context of content moderation.

2.1. Defining transparency: a four-layer framework

In its most basic definition, derived from physics, transparency refers to "the characteristic of being easy to see through" (Transparency, n.d.). Influenced by the overarching Freedom of Information laws from the 1980s (Meijer, 2014) and the misleading sentiment that seeing is knowing (Ananny and Crawford, 2018), *seeing through* has come to be equated with *information disclosure* in popular discourse. However, scholars unanimously contend that mere information disclosure, or openness, is a simplistic conceptualisation of transparency that overlooks the necessity for this information to be received (Kosters and Gstrein, 2024; Ananny and Crawford, 2018) and leveraged (Suzor et al., 2019). Just because an organisation publishes data does not mean that it is transparent. This conceptualisation is the equivalent of Fox's (2007) *opaque transparency* or Heald's (2006) *nominal transparency,* that is, public but empty information that does not reveal how institutions behave in practice, because the data is too complex, irrelevant and/or unreliable. It also echoes the normative ideal of *transparency as a virtue* that values openness but fails to specify what it relates to in practice (Meijer, 2014). This conception confuses transparency with knowledge itself, while it should only be considered a facilitator (Schauer, 2011). To increase in meaningfulness, a few scholars mention that transparency could comprise *publicity* of information (Naurin, 2007), that is, data availability should be brought to people's attention (Forssbæck and Oxelheim, 2014). Most importantly, the literature emphasises that meaningfully transparent information should be *understandable to an audience.* This comprises two elements. First, meaningful transparency recognises that transparency is an institutional relation that links an emitter and a receptor (Meijer, 2014). Second, it implies that at least a part of this audience is able to process the information and leverage it to undertake some action. Kosters and Gstrein (2024) call this *clarity of information*, while Heald (2006) calls it *effective transparency*. For Forssbæck and Oxelheim (2014), in economics, the very definition of transparency is the reduction of information asymmetries, which implies understandability. Eminently, Rieder and Hofmann (2020) have coined the term *observability* to refer to the ability to collaboratively produce knowledge on the emitter of transparency (in their case, a digital platform) and fundamentally understand what its impact on social relations and society is at large. Finally, some scholars believe that the information should be *audience-tailored* (Suzor et al., 2019; Douek, 2022), that is, presented and filtered differently according to the several groups constitutive of the audience (Kosters and Gstrein, 2024).

From these links across the literature, it is possible to create a four-layer framework to conceptualise transparency (see Figure 1), which extends Kosters and Gstrein's (2024) observation that transparency has three layers. Among these layers, two -disclosure and understandability- are especially important, the others being less extensively discussed in the literature. All four levels are different definitions of transparency that build on the ones below them. For example, transparency as publicity includes transparency as disclosure. This polysemy can trigger semantic uncertainty: some scholars value "transparency" because they define it as understandability (Forssbæck and Oxelheim, 2014), while others reject "transparency" because they define it as information disclosure (Rieder and Hofmann, 2020). To clarify, this thesis uses the concept of *meaningful transparency*, an existing term in the literature that is unevenly mobilised, to refer to the framework's levels 3 and above.

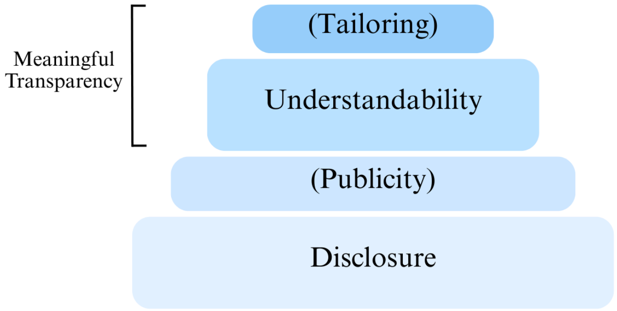


Figure 1: The four definitions of transparency

In order to fine-tune this framework, it is important to include the what, who, to whom, when and how questions. They emphasise that transparency is not binary but can rather exist in varying degrees and natures (Schauer, 2011). Further, they are essential components of the definition of transparency: at the very least, transparency supposes an emitter, a receptor and an information to be transmitted (Schauer, 2011). The first question is, *what* is made transparent? In general, Heald (2006) distinguishes between event transparency and process transparency. Events are punctual outcomes, while processes refer to how these outcomes have been reached. Second, *who* is transparent? Transparency has historically concerned state entities but it has spread over the private sector in regulated areas and more recently over the tech sector. Within the same organisation, transparency can focus on individuals or the institution as a whole (Fox, 2007). Third, *to whom* should this actor be transparent? Heald (2006) distinguishes between horizontal and vertical transparency, depending on whether there is a hierarchical relationship between the actor and the targeted forum (Söderlund et al., 2024). Fourth, *when* is the information made transparent? Though it is usually ex post, it can also be continuous (Heald, 2006) or even ex ante when focusing on intentions (Forssbæck and Oxelheim, 2014). Finally, *how* is the information made transparent? Release can be proactive, demand-driven (voluntarily or not) or forced (by whistleblowers, leaks or legal requirements). The next part will focus on operationalising this framework in the context of content moderation.

Before that, it is crucial to mention how transparency differs from, and perhaps how it bridges with, *accountability*. Accountability refers to "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences" (Bovens, 2007, p.450). Though they are closely related, transparency and accountability are two distinct concepts. In particular, accountability systematically implies scrutiny by a specific forum and the possibility of sanction (Bovens, 2007) while it is not the case in all conceptions of transparency. How are the two connected? For Fox (2007), there exists "soft accountability" which refers to the capacity for a forum to demand answers without being able to translate this judgement into tangible consequences. In that sense, meaningful transparency equates to soft accountability, but it unsatisfactorily supposes that the possibility of sanction is not a defining feature of accountability. Within our definition, meaningful transparency can be said to *foster* (Suzor et al., 2019), or even be a *prerequisite* for (Meijer, 2014; Bovens, 2007), accountability. It is easy to see how available understandable data can help reveal malfunctioning or reward success. Yet this relation is not systematic, and might even lead to opposite results: some actors can present increased transparency as a substitute for accountability (Meijer, 2014).

2.2. Defining meaningful transparency in the context of content moderation on social media platforms

To get around the limitations of transparency discussed in section 1.3., proponents of transparency measures for platforms emphasise that effective transparency should be *meaningful*, in the sense of the above framework. However, there is a lack of operationalisation as to what providing meaningful transparency on content moderation means in practice. The literature either cites meaningfulness as a general goal without specifying its meaning (Angus et al., 2023; Reid and Ringel, 2025), produces tautologies (meaningful transparency means providing meaningful information) (Kaushal et al., 2024) or sporadically names practical elements that platforms should fulfill without, to the author's knowledge, combining these contributions into a thorough common understanding. To assess whether social media platforms provide meaningful transparency, we need to go beyond the simple definition of transparency as understandability and ask the aforementioned questions: who understands whom? what, when, how? Using the literature on content moderation transparency, the following tries to respond to these questions to operationalise meaningful transparency in the context of content moderation on social media.

Starting with the directions of transparency (who, to whom), the emitter of transparency can be governments, regarding their requests (Citron, 2018), or the platform, at organisational or personal level. Then, platforms can be transparent *to* many different groups, namely researchers, civil society, regulators, users, the public or their environment. Depending on the provision, one or several of these actors can be included in the audience. Still, the reviewed literature is unanimous to say that there needs to be an effective oversight for content moderation transparency to be meaningful (Gorwa and Ash, 2020). In this sense, regulators are necessarily included in the audience, alone or backed by independent auditors (Douek, 2022), but there is also a consensus on the necessity for researchers to hold platforms accountable, especially through an access to platform data (Suzor et al., 2019; Lewandowsky, 2024). A part of the literature underlines that civil society actors too should have access to this data (Suzor et al., 2019), as they are essential translators of this information to the relevant publics (Reid et al., 2024a). However, if the data is not publicly available, it can be harder for them to get vetted access because of the lack of formal membership and the high turnover that characterise the sector (Leerssen, 2020). Precisely, there are debates on how much information should be available to the general public or kept restricted. Arguments against open access include the necessity to protect trade secrets, data privacy (Bassan, 2025), but also the probable lack of interest from most end-users. However, if the goal is to lower the discretionary power of platforms, there might be a case for of multistakeholder engagement (Helberger et al., 2018), including for reasons of legitimacy. For Bassan (2025), an absence of information should signal that platforms have a reason not to disclose it, hence opening access would show that nothing is being hidden. Public access can satisfy privacy concerns if done appropriately, and serve as a starting point for subsequent in-depth research relying on more sensitive data (Leerssen, 2020).

Next, the operationalisation of meaningful transparency should specify *what* should be made transparent. We can group the sporadic inputs from the literature into three categories. The first one contends that users should be able to understand the moderation actions that are imposed on their content. It is close to the digital constitutionalism project and its emphasis on procedural justice. This category encompasses three elements. First, users should be informed about the fact that one of their pieces of content has been actioned. The notice should detail the exact nature of the offending content (Suzor et al., 2019; Bassan, 2025), through a direct URL or a description, and be permanently available to the user (Suzor et al., 2019). Second, platforms should provide "detailed and individualized" explanations for the sanction, addressing why this decision was met and how the content was flagged for review (Suzor et al., 2019; De Gregorio, 2020), using local explanations and natural language (Palmeira Ferraz et al., 2024). This implies that the platform's terms and conditions should be publicly accessible (Castets-Renard, 2020). Plus, if they don't agree, users should be able to appeal through a clear and easily accessible procedure (De Gregorio, 2020; Bassan, 2025). Third, anyone should be able to understand how content moderation is enforced. This supposes "deep qualitative analyses of the automated and human processes that platforms deploy internally" (Suzor et al., 2019, p.1538), including information about training and guidelines given to moderators, and algorithmic and AI transparency. A specific literature deals with AI transparency, outlining a number of criteria that these machine learning algorithms should fulfill to be understandable, from explaining the training process to how it works in practice. This includes disclosing what data is used and how it is being processed, explaining decision-making criteria and their justifiability and ensuring auditability (Felzmann et al., 2020), inclusivity, interventability and interoperability (Yu, 2021). Enforcement transparency also needs to account for failure (Gorwa and Ash, 2020), by measuring its frequency, its nature, and being open about the processes implemented to deal with mistakes and novelties (Suzor et al., 2019; Felzmann et al., 2020). The second category of transparency objects is the need for anyone interested (or at least researchers, civil society actors and regulators) to understand moderation sanctions at aggregate level. Apprehending the moderation of platforms cannot be done by merely looking at individual sanctions: having a sense of what the sum of these looks like is necessary. This translates in large-scale access to data on moderation decisions (Suzor et al., 2019) and information provision on specific and consistent metrics (Reid and Ringel, 2025), often in the form of a report, for example on the number of content actioned within a period, their nature and their modalities (a list of specific data points can be found in the Santa Clara Principles on Transparency and Accountability in Content Moderation (2018)). This report should be accessible, usable, and easily understandable (Bassan, 2025). Finally, the third and last category is about understanding moderation at systemic level. This part of the literature takes a broader look by considering moderation not only through sanctioning but rather through its larger meaning of sorting and acting on content, as was defined in section 1.2. Thus, advertising and recommender system transparency belong to this category. On advertising, users should be able to understand when and why they are presented an ad and whether it is targeted (Jost et al., 2023), while on recommender systems they should be provided with non-technical and straightforward (Wang, 2024; Haque, 2024) explanations on the algorithm in a visual and interactive way (Luria, 2022) along with ways to increase user controls (Storms, 2022; Mitova et al., 2023). Papers from this category invoke a more quintessential understanding of moderation practices as they demand to be able to understand the social impacts of moderation (Bassan, 2025; Suzor et al., 2019) and how digital platforms affect user behavior (Bassan, 2025), society (Rieder and Hofmann, 2020) and public discourse (Bassan, 2025). For that purpose, the whole hierarchy of moderation should be known (Douek, 2022), including how external and internal decision-makers operate (Suzor et al 2019; Bassan 2025).

Finally, the modalities through which transparency is produced also matter in defining meaningfulness. In terms of timing (when), the most widespread practice is that of regular ex post accounts of the actions performed by platforms within a given time period, but Rieder and Hofmann (2020) argue that the most meaningful observation of platforms is continuous, because of how volatile they are. Some also make the case for ex ante transparency (De Gregorio, 2020), notably through notices of how specific issues *will be* handled or how moderation will be done, but it is not enough to monitor platforms' doings. Looking at how things should be made transparent, there is an understanding that voluntary transparency has several times proved unhelpful (Portraru, 2017; Leerssen, 2020), because it reveals only what portrays the platform positively. The most meaningful information would then originate in forced access. Because it is undesirable to rely on whistleblowers to provide transparency to the democratic forum, legislation could be best placed to foster meaningful transparency.

Figure 2 summarises the operationalisation of meaningful transparency that this section has undergone. The central component concerns *what* is made transparent. The "who" and the "to whom" are often implied in the level itself, and the "when" and the "how" are modalities. Because not all authors mention all categories in their apprehension of meaningful transparency, a platform can be meaningfully transparent at one level but not at another one. This can now be used to assess meaningful transparency in practice.

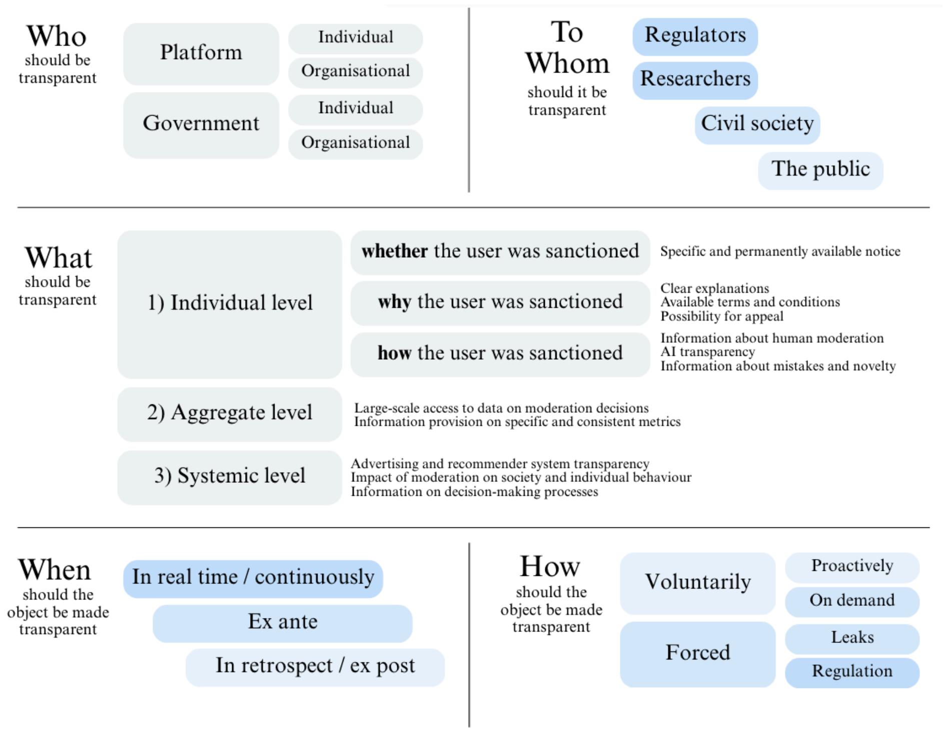

Who
should be transparent
Platform
Individual
Organisational
Government
Individual
Organisational
To Whom
should it be transparent
Regulators
Researchers
Civil society
The public
What
should be transparent
1) Individual level
whether the user was sanctioned
Specific and permanently available notice
why the user was sanctioned
Clear explanations
Available terms and conditions
Possibility for appeal
how the user was sanctioned
Information about human moderation
AI transparency
Information about mistakes and novelty
2) Aggregate level
Large-scale access to data on moderation decisions
Information provision on specific and consistent metrics
3) Systemic level
Advertising and recommender system transparency
Impact of moderation on society and individual behaviour
Information on decision-making processes
When
should the object be made transparent
In real time / continuously
Ex ante
In retrospect / ex post
How
should the object be made transparent
Voluntarily
Proactively
On demand
Forced
Leaks
Regulation

Figure 2: Summary of what meaningful transparency means in content moderation
The shade of blue indicates the degree of meaningfulness when it is identifiable.

3. The DSA: introducing transparency to mitigate the undue power of platforms

Since the mid 2010s, there has been a sense of loss of agency and power over the internet in the EU, at individual and union level (Savin and Trzaskowski, 2023). With platforms getting increasing market and social power (Flew and Gillett, 2020), EU tech companies lagging behind (Savin and Trzaskowski, 2023), technology further entrenching in daily lives and at a younger age (Kabali et al., 2014), fears over disinformation and hybrid warfare getting prestance and AI technology developing at a fast pace, the landscape now looks very different than it did in the early 2000s. Far from the early narratives of the internet around freedom and escape from (government) regulation (Lessig, 1995), we are now witnessing a regulatory turn over the internet. Several massive texts in their scope, namely the GDPR (2016), the sister regulations DMA (Digital Markets Act) and DSA (2022) and the AI Act (2024), were adopted over the last ten years in the EU.

Proposed in late 2020, the DMA (Regulation 2022/1925) and DSA (Regulation 2022/2065) were adopted in September and October 2022, after a remarkably fast legislative process. While the DMA legislates the economic aspect of platform dominance by addressing abuses of dominant position and mandating interoperability, the DSA regulates the relationship between online service providers and the information that they host, transmit, and capitalise on. In particular, it puts the emphasis on *transparency*. This section first introduces the legislative background behind the DSA and proceeds by discussing the substance of the DSA through the lens of transparency.

3.1. Background: the need for a modernised framework to regulate digital services

Up until today, the central piece of legislation regulating digital platforms has been the e-commerce Directive (2000/31/EC), drafted between 1998 and 2000, at a time when neither touch phones nor Facebook existed. This directive lays down harmonised rules on the whole life cycle of electronic commerce activities, i.e. the establishment of information society service providers in the EU, commercial communications (including advertising), online conclusion of contracts and liability of intermediaries (Lodder, 2017). It also introduces a three-pole typology of intermediary services providers: mere conduit services (which merely transmit information), caching services (which temporarily store information to transmit it more efficiently) and hosting services (which store information). Crucially, it articulates three principles that have since constituted the foundation of EU tech regulation (Madiega, 2022): 1) the country of origin principle (i.e. information societies are subject to the laws of the country that they are established in), 2) the limited liability regime (i.e. online intermediaries are not responsible for the content that they host, under certain conditions), 3) the no-general-obligation-to-monitor principle (i.e. platforms are not required to police the content that they host). The goal of such a regime was to create incentives for innovation and investment in the nascent internet market, still permeated by the idea of a free space (Frosio, 2023).

While the e-commerce Directive has worked relatively well for twenty years (Schwemer, 2023), some provisions have become outdated given the evolution of the digital landscape. In particular, its drafters had not foreseen the extent to which digital services would develop, nor had they anticipated the privatisation of enforcement and the subsequent oversight challenges that it would entail (Schwemer, 2023). Besides, it became increasingly salient that internet communications could amplify and accelerate the spread of hate speech and misinformation. Given their extent and, according to Citron (2018), in reaction to the 2015 terrorist attacks, some Member States chose to create their own legal framework against online harms. In particular, Germany adopted the NetzDG (2017), France the Loi Avia (2020) and Austria the KoPl-G (2021). In addition to facing constitutional and/or European law conformity challenges, these laws fragmented the internal market by creating legal uncertainty for intermediaries (Buri and van Hoboken, 2021; Recital §2 DSA). This contradicted the opportunity that the (nearly) borderless internet provided for building a true digital Single Market (Spindler, 2023), which was one of Jean-Claude Juncker's core strategies during his mandate as President of the European Commission (2014-2019). For its part, the EU had taken on a soft-law approach to regulating online harms, by enabling voluntary self-regulation with a handful of large online platforms, notably through the 2016 EU Code of Conduct on Countering Illegal Hate Speech Online and the 2018 EU Code of Practice on Disinformation. These reflected a change of rhetoric in the perception of platforms' responsibility and need for accountability, compared to pre-2016, but not a change of policy (Farrand, 2023). While the Commission praised their efficiency, including by emphasising swift removal rates (European Commission, 2019), some scholars worried that such approaches would lead to censorship (Portraru, 2017; Citron, 2018).

3.2. A tiered regulation centered around rights and transparency

The DSA seeks to improve the functioning of the internal market by creating a "safe, predictable and trusted online environment" (Article 1 DSA) which protects innovation and fundamental rights. Beside enforcement measures, the DSA's provisions can be grouped in two parts. First, the regulation specifies rules pertaining to the (non-)liability of providers of intermediary services. In effect, it takes on almost verbatim the core provisions of the e-commerce Directive, especially the liability exemptions (Schwemer, 2023). Under this framework, hosting services are exempt from liability as long as they do not have knowledge of illegal content, or, when they do, act "expeditiously to remove or disable access to the illegal content" (Article 6(1)(b) DSA). When analysing social media platforms, as this thesis does, hosting services are the most relevant category to take into account because they include online platforms and their subset, the VLOPs, two additional layers of detail added by the DSA onto the e-Commerce directive's typology (see Figure 3).

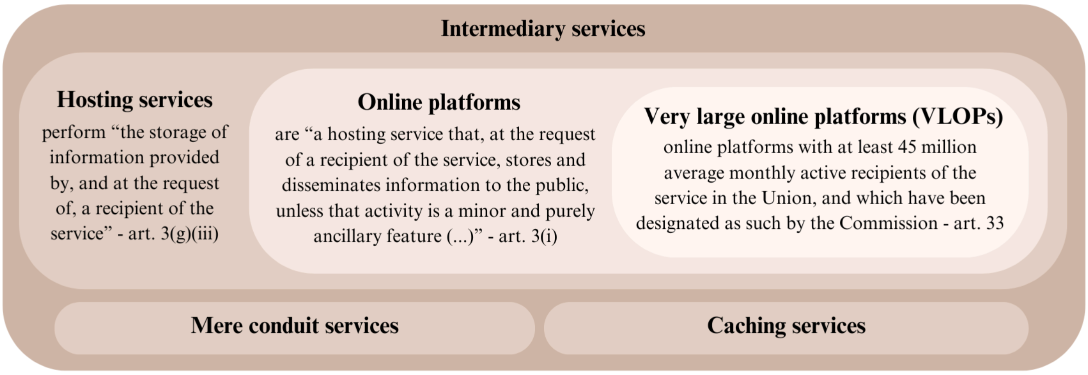
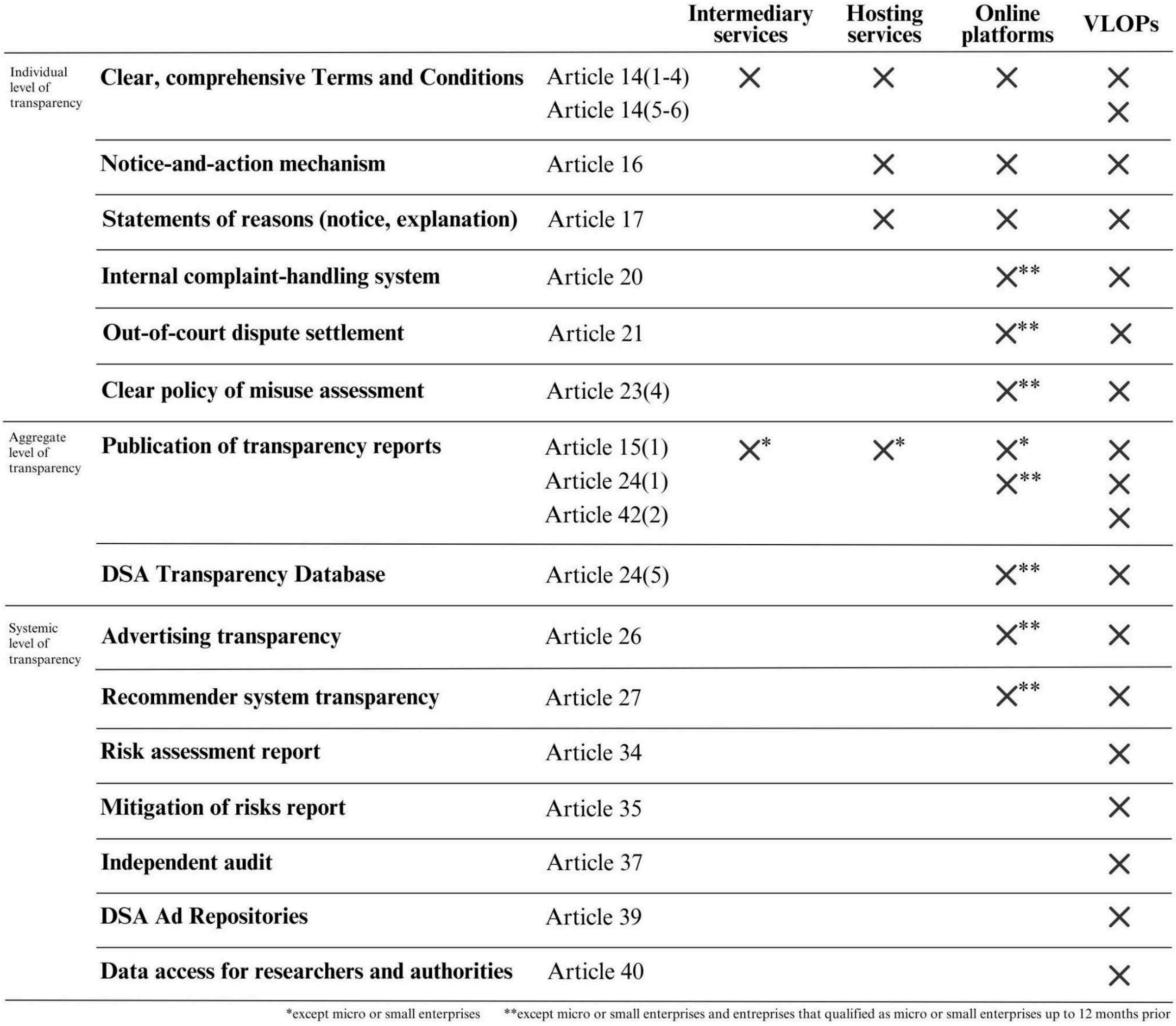

Intermediary services
Hosting services
perform "the storage of information provided by, and at the request of, a recipient of the service" - art. 3(g)(iii)
Online platforms
are "a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature (...)" - art. 3(i)
Very large online platforms (VLOPs)
online platforms with at least 45 million average monthly active recipients of the service in the Union, and which have been designated as such by the Commission - art. 33
Mere conduit services
Caching services

Figure 3: Types of intermediary service providers under the DSA

Second, the DSA creates due diligence obligations applying to intermediary services providers, with the goal of increasing providers' accountability. These obligations broadly center around content moderation and systemic risks mitigation. The DSA takes on a risk-based approach by imposing increased burdens where risks are the highest, namely to very large online platforms (VLOPs) and search engines (VLOSEs), whose user base represents at least 10% of the EU population. It is also mindful of compliance costs, which VLOPs/VLOSEs are best equipped to handle. As of April 2025, 23 platforms have been designated as VLOPs, ranging from Meta to Zalando, and 2 as VLOSEs (European Commission, 2025). In case of suspected non-compliance, the Commission, which is tasked with overseeing VLOPs/VLOSEs, investigates before potentially opening official proceedings. Sanctions include fines up to 6% of the worldwide annual turnover, periodic penalties (Article 52 DSA), and in last resort a temporary suspension of service, issued by a judge (Article 51(3)(b) DSA). A summary of the main DSA provisions is presented in Figure 4, along with the providers to which they apply. It can be observed that sorted in order of appearance, they match with the three substantive conceptualisations of meaningful transparency previously introduced (the "what" question), which are written in the left column. Provisions will be discussed in the analysis, which makes Figure 4 the roadmap of this thesis.



			Intermediary services	Hosting services	Online platforms	VLOPs
Individual level of transparency	Clear, comprehensive Terms and Conditions	Article 14(1-4)	✕	✕	✕	✕
		Article 14(5-6)				✕
	Notice-and-action mechanism	Article 16		✕	✕	✕
	Statements of reasons (notice, explanation)	Article 17		✕	✕	✕
	Internal complaint-handling system	Article 20			✕**	✕
	Out-of-court dispute settlement	Article 21			✕**	✕
	Clear policy of misuse assessment	Article 23(4)			✕**	✕
Aggregate level of transparency	Publication of transparency reports	Article 15(1)	✕*	✕*	✕*	✕
		Article 24(1)			✕**	✕
		Article 42(2)				✕
	DSA Transparency Database	Article 24(5)			✕**	✕
Systemic level of transparency	Advertising transparency	Article 26			✕**	✕
	Recommender system transparency	Article 27			✕**	✕
	Risk assessment report	Article 34				✕
	Mitigation of risks report	Article 35				✕
	Independent audit	Article 37				✕
	DSA Ad Repositories	Article 39				✕
	Data access for researchers and authorities	Article 40				✕

*except micro or small enterprises **except micro or small enterprises and entreprises that qualified as micro or small enterprises up to 12 months prior

Figure 4: (main) Transparency provisions expected from intermediary service providers under the DSA

The DSA is a unique legislation that might be setting the basis of a new generation of rules for internet intermediaries (Husovec, 2023). To counter asymmetry, it empowers users (Husovec, 2023), researchers (Söderlund et al., 2024), and potentially civil society at large (Eder, 2024), with new rights, missions and information. By doing so, it seeks to rebalance platform governance so that everyone in the democratic forum, including providers, effectively does their part, without blaming providers for all ills (Husovec, 2023). Transparency and proceduralisation are the core mechanisms serving this goal, which critical transparency scholars, however, see as forces legitimising the undue power of platforms (Maroni, 2023). Other novelties brought by the DSA are the shifts from self-regulation to mandated regulation and from entirely ex post to partially ex ante transparency (Turillazzi et al., 2023). Finally, it should be noted that the DSA refrains from setting substantial requirements, i.e. it does not mandate providers to delete a specific type of content, but rather asks providers to create their own rules and enforce them fairly (Husovec, 2024). Given the directive's innovative nature, it is crucial to scrutinise how the DSA changes the practices of providers and the balance of power in practice, even more so because it is likely to have influence abroad and translate into a “Brussels effect” (Nunziato, 2023).

While the DSA’s substance (Husovec, 2023; Turillazzi et al., 2023) and potential effects on transparency (Maroni, 2023; Söderlund et al., 2024) have now been discussed extensively, focusing on particular aspects like disinformation (Nannini et al., 2024; Husovec, 2024), systemic risks (Palumbo, 2024), advertising (Izyumenko et al., 2024; Duivenvoorde and Goanta, 2023) and ranking (Leerssen, 2023), or on particular cases like TikTok (Kosters and Gstrein, 2024), there are yet few studies that investigate the DSA’s *observed* impact on transparency. Some have discussed it in relation to the first DSA jurisprudence (Van De Kerkhof and Goanta, 2024) and others in relation to the nascent Transparency Database (Trujillo et al., 2024; Kaushal et al., 2024). Yet to the author’s best knowledge, no study has yet tried to observe the DSA’s *overall* impact in situ, to the extent that most of the material used in this thesis is widely understudied. Given the scope of such a scheme, this thesis is only a first step towards bridging this gap.

4. Meta’s historical approach to content moderation and transparency

This thesis considers the particular case of Meta (“Facebook” before 2021), the company that owns Facebook, Instagram, WhatsApp, Messenger and Threads. Run and founded by Mark Zuckerberg, who owns 13,5% of the company and holds 61% of the voting power (Reiff, 2024), Meta is one of the United States' tech giants. Each of its three main products (Facebook, Instagram, WhatsApp) has between two and three billion monthly active users worldwide (Statista, 2025) and the company’s net income was more than 62 billion US$ in 2024 (Meta, 2025a). Before we reflect in the next section on the methodological arguments and implications behind the choice of this case study, it is important to briefly describe what Meta is and how it has historically positioned itself regarding the issues at stake in this thesis, namely content moderation and transparency. This section provides the necessary context to be able to assess the impact of the DSA on Meta.

Today, transparency (or openness) is among the values that Meta considers the core of its identity. This is reflected in its external communication, for example when it emphasises that it is an industry leader on transparency or that it fosters more “open and honest communities” (Gorwa and Ash, 2020), but it can also be found in its internal corporate culture: Meta’s career website highlights its unmatched “culture of collaboration and transparency” (Meta, n.d.a) and its headquarters at Menlo Park (US, California) are a gigantic openspace with glass-walled meeting rooms and open desks, including for its CEO Mark Zuckerberg (Gorwa and Ash, 2020). However, this contrasts with the reality of Meta’s governance practices. Facebook has been described as opaque by ex-employees-now-whistleblowers Frances Haugen (Ouangari, 2021) and Sarah Wynn-Williams (Urwin, 2025), but also by researchers (e.g. Williams (2023)) and NGOs. For example, nondisclosure agreements are frequent (Hern, 2021; Perrigo, 2022), if not systematic, to lower the risk of whistleblowing. As a result, little is known of what happens at Meta and society increasingly depends on the said whistleblowers to get this information (Olesen, 2025).

Precisely, this points at the primary characteristic of Meta's relationship with transparency on content moderation: historically, transparency has been almost exclusively *forced*, resulting from leaks, or *reactive*, answering to societal pressure after various scandals. The first act of transparency that Facebook implemented was in 2013, when it issued its first transparency report in reaction to the Snowden case (Reid et al., 2024a). At the time, transparency was regarded as a potential safeguard against abusive government censorship, therefore the report focused exclusively on government requests addressed to Facebook. The 2016 elections marked the first "scandal" directly questioning how Facebook manages the risks stemming from its platform. Donald Trump's election as President of the United States brought unprecedented attention to the spread of misinformation on social media, but more broadly to the virality that social media allows. Plus, investigations later revealed that Russia had waged a coordinated disinformation campaign on Facebook and other social networks to interfere in the election (Frenkel and Bennel, 2018). In response, Facebook underwent design updates, for example increasing the transparency of political advertising (Jost et al., 2023), and joined the EU Code of Conduct on Countering Hate Speech Online (2016). In 2017, Facebook faced an extensive crisis of trust: former executives were speaking out against the company, stating that it was putting profit ahead of users' well-being (Newton, 2017), and studies were denouncing social media use as being detrimental to mental health, which the company acknowledged at the end of the year (Levin, 2017). Mark Zuckerberg pledged to "fix Facebook in 2018", but the platform remained relatively untouched (Karanicolas, 2021) and (hence) widely opaque. 2018 was a substantial gamechanger: with the Cambridge Analytica privacy scandal, Facebook's share lost 19% in one day (Neate, 2018), judicial inquiries got opened in the US and the UK, and a global boycott campaign #DeleteFacebook gained momentum among users, advertisers and prominent tech actors like WhatsApp co-founder Brian Acton (Gerken, 2018). As a response, Facebook released its Community Standards in April, launched an ad archive, announced the creation of an independent Oversight Board, underwent numerous design updates and reinforced its self-regulation commitments by joining the EU Code of Practice on Disinformation (2018) and the Santa Clara Principles on Transparency and Accountability in Content Moderation (2018). During this period, Facebook bolstered its "proactive" transparency (i.e. not mandated by law, but reactive in essence) to show its diligence and regain the trust of users, investors and legislators alike. Subsequently, Facebook's actions against health-related misinformation during the Covid-19 pandemic (2020) and the ban of Donald Trump from the platform after the US Capitol Riots (January 2021) ended to achieve the involvement of platforms in content curation (Karanicolas, 2021). At this moment launched the main governance mechanisms that Meta relies on today to promote its transparency: first, the Transparency Center (May 2021), a dedicated website that centralises Meta's various transparency reports, policies (including Community Standards) and explanations on the functioning of its services; second the Oversight Board (October 2020). This Board, often nicknamed "the Supreme Court of Facebook", is a unique structure: an irrevocable trust endowed with $130 million in 2019 and a further $150 million in 2022 that grants users the possibility to ask its twenty members (eminent academics, journalists, lawyers, politicians) to review their case after their appeal within the app was unsuccessful. The Board's stated usefulness is to bring "transparency to content moderation processes that were often unknown or unclear" (Oversight Board, n.d.), and it itself publishes transparency reports (though none seem to have been published since 2023). In practice, the Oversight Board is undeniably a legitimacy-building investment (Price and Price, 2023). Whether it indeed increases transparency is debated: some praise the opening of social media decisions to independent oversight (Ang and Haristya, 2024) but others underline how focusing on individual decisions distracts from questioning Meta's overall policies and power (Douek, 2022), especially given that the Oversight Board does not have access to any internal data (Oversight Board, 2021).

Whether it is effective or not does not change the fact that since 2020 Meta has been *positioning* itself as the industry leader on transparency, and also as rather responsive to regulation. Still, it is worth mentioning that seven requests for information were sent to Meta on Facebook and/or Instagram under the DSA, followed by two official proceedings in April and May 2024, which are still ongoing as of April 2025. Further, Meta's positioning could change after Mark Zuckerberg's recent rapprochement with Donald Trump, who holds that the EU uses lawfare against US tech companies, and the policy changes and comments that followed (e.g. Mark Zuckerberg said that Europe was "institutionalising censorship") (Rankin, 2025). As Gillespie (2018) puts it, the values behind social media's policies are not mostly some moral core, but the best compromise possible between competing pressures.

5. Methodology, data and sources

To answer the research question, this thesis uses a qualitative case study. Qualitative research methods study their object of research holistically, "specifically to unravel a complex phenomenon or one with little information about" (Njie and Asimiran, 2014, p.35). As was outlined, transparency is a complex phenomenon *and* when it is not achieved, which this study tries to investigate, little information is available on the actor/organisation in question. Qualitative methods are therefore highly relevant to analyse this interactive process that cannot be captured by figures. The choice of conducting a case study flows naturally. Social media are far from being a monolith and hence require detailed and individualised inquiries.

The case study under investigation is Meta, the company described in the previous section. There are three characteristics in Meta that make it an interesting case. First, Meta has an extremely large volume of users and is one of the biggest companies in the social media industry. This implies three things. It means that its content moderation policies have a worldwide impact and that each individual user is in a position of extreme asymmetry regarding the company, which makes the need for transparency and the need for assessing this transparency greater. This further means that even though the DSA does not target any company, if the legislation is effective in attaining its stated goal of making large social media platforms more transparent, then it should increase Meta's transparency in a measurable and significant manner. The final interesting element about Meta's size is the fact that the same company owns several products subjected to the DSA, two VLOPs (probably soon three, as WhatsApp just announced that it crossed the 45M users threshold) and smaller online platforms, which makes it theoretically possible to investigate services of different sizes while opening up questions of consistency and differentiation across the products of a single company. The second characteristic that makes Meta worth studying for this research question is the fact that the company has launched a number of voluntary initiatives to increase transparency in the past years, as described in the previous section, which Meta is very vocal about. This suggests that more data might be available than for other companies, or at least that it goes back further in time, which is useful to assess the extent to which the DSA *increased* transparency. Finally, the third interesting characteristic, which goes hand in hand with the second one, is that Meta has faced numerous challenges in the recent past on how it manages the risks inherent to its platform and on how little transparency it actually provides. This, in itself, prompts us to know whether a legislative tool such as the DSA can be effective in increasing a famously opaque system. In addition, the contrast with the previous characteristic also exhorts us to scientifically assess Meta's actual level of transparency to challenge the company's discursive power. Though this case is often selected in the literature, there is to the author's best knowledge no study that investigates the overall impact of the DSA on Meta Platforms.

Several types of data are used to conduct this research. All of them are collected through desk research to reflect the rationale behind this investigation: assessing objective transparency, i.e. "the extent to which systems release information about how they work" (Guesmi et al., 2023, p.3), and the extent to which this information is available to any user, researcher or civil society actor based in the EU. Throughout the analysis, though to a varying degree, three types of written documents produced by Meta are used. The first one are communications, blog posts and press releases found on Meta's and Facebook's newsrooms. They are not collected systematically but thematically, when they deal with topics relevant to the present investigation, over the period 2018 until today. The second category includes transparency "tools" designed by Meta: the Ad Library, the Transparency Center and the Help Centers. Driven by the necessity to investigate what Facebook's transparency looked like in the past, the Transparency Center is browsed in its current version but also using the Internet Archive / WayBack Machine to go back as far as 2011, on the ancestor URLs of the Transparency Center (www.facebook.com/communitystandards, https://transparency.facebook.com, https://transparency.fb.com). Finally, the various written documents produced by Meta under the DSA are investigated, including the risk assessment reports (Meta, 2024a; Meta, 2024b), transparency reports (Meta, 2023a; Meta, 2024c; Meta 2024d; Meta, 2024e), and to a smaller extent the independent audit report (EY, 2024). Throughout the analysis, newspaper articles are also used, mostly from the specialised press on tech and social media.

To assess user-facing transparency, I use screenshots of Facebook and Instagram collected from my own browsing and from two acquaintances based in the EU who were sanctioned in the studied timeframe. On Instagram, I also report two inappropriate comments and post on a dedicated account in March 2025 one picture that I am aware to be against the Terms of Use in order to see the user-facing statement of reason provided by Instagram upon deletion. On the Transparency Center, Meta only shows a part of the user-facing notice, so posting a picture was the only way to see the entirety of the notice. Appropriate safeguards are taken to minimise the ethical challenges stemming from this. The picture is under GNU Free Documentation license and is not obscene, as it is an educational picture of female breasts. The picture is posted on Instagram only, which is not under real-name policy contrary to

Facebook. The account used has no followers and the picture stayed up online for less than two minutes, so that it is quasi certain that no one has seen the picture. No appeal was lodged.

To study the aggregate level of meaningful transparency, which refers to the sum of individual sanctions and hence implies some level of quantitative data, I investigate the Transparency Database (TDB) (European Commission - DG CONNECT, 2023), which is an emblematic yet understudied provision of the DSA. A total of 635,6 million of statements of reasons (SoR) logged between April 01, 2024 and March 12, 2025 are downloaded from the TDB's website, coming in daily zipped CSV files, for Facebook, Instagram, Threads and WhatsApp Channels. VLOPs must comply since the launch of the TDB on September 26, 2023, while smaller platforms theoretically must since February 2024 but got more leeway: Threads started to log statements on September 05, 2024 and WhatsApp Channels on September 17, 2024. Therefore the entirety of statements is collected for these two platforms until March 12, while for Facebook and Instagram the choice is made to shorten collection from April 01 onwards, for two main reasons. First, I choose to limit the analysed time period to ease material feasibility, because Facebook in particular has a lot of SoR and it can be difficult to handle for a personal computer's working memory. Second, the specific date of April 01 is selected to be able to compare the data from the TDB with the data from Instagram and Facebook's last transparency reports, whose reporting period ranges from April 01 to September 30, 2024. The remaining time period is of eleven months, which satisfies the goal of looking at the evolution of moderation patterns through time, and deals with rather uninvestigated data, since four out of the five existing studies on the TDB investigate SoR from 2023. These four platforms are selected because they were the only Meta platforms available on the TDB at the time, with the addition of "Meta Horizons" but the latter was excluded because it had logged zero SoR and does not correspond to this research's object of inquiry, i.e. social media platforms. Finally, I choose to download the "full" version of the SoR to be able to investigate the free text justifications offered by Meta and the territorial scope of application of its decisions, which are not included in the "light" version of the files.

The study is structured along the three levels of meaningful transparency identified in the theoretical framework (individual, aggregate and systemic levels). For each level of meaningful transparency, five questions are asked:

(Q1) What provisions in the DSA aim at increasing transparency at this level?

(Q2) What voluntary practices existed before the DSA was introduced (Dec 2020) that are similar to the provisions identified in (Q1)?

(Q3) Has there been an increase in transparency between (Q2) and today?

(Q4) If an increase in transparency is observed, has it been triggered by the DSA? (If it is possible to tell)

(Q5) How meaningful is this DSA increase in transparency? i.e./or How meaningfully transparent is Meta today?

For each category, the goal is therefore to assess whether Meta goes beyond information disclosure (i.e. meaningful transparency) and if so, how far. When addressing specific products and not only Meta, it especially focuses on Instagram and Facebook because they are Meta's biggest platforms and also those that moderate the most. Threads is still rather new and is highly similar to Instagram since it is half embedded in it. WhatsApp Channels is not based on the same model: it moderates very little, has no ad and few recommender systems, so that it is less relevant. Furthermore, "Meta's past practices" in (Q2) is most of the time taken as a synonym of "Facebook's past practices" because "Meta" was previously called "Facebook" and often applied its policies first to Facebook, then to its other products.

To analyse the TDB in particular, I use Python's "Polars" package to treat and investigate the data, and Python's "Matplotlib" library to plot the data. In line with principles of Open Science, the code is available in the following public repository: https://github.com/lolapttr/Master_thesis_TDB. To conduct the analysis, the 635,6 million SoR logged between April 01 and March 12 are narrowed down to 625,2 million. Indeed if we look at all the SoR logged by a Meta platform on a given day, these SoR do not solely include moderation actions that were performed on that specific day. When a moderation action is performed, it usually takes between one and two days for the SoR to be logged onto the database (for 60-76% of SoR), but it can go up to six days (seven and eight days for Threads and Instagram, but marginally). Hence, to analyse comprehensively the content actioned over a given time period, it is necessary to collect the logs six days further than the last date of the sample, but considering only the logs corresponding to that last day. To illustrate, this means that to analyse sanctions *performed* until March 06, 2025, I collect the SoR *logged* until March 12, 2025 and within this six-day difference, I only keep the sanctions performed on March 06. This way, I include (almost) all the moderation actions performed on March 06 even if they were logged six days later. For the same reasons, I conversely filter the logs of the first days of my sample to remove the SoR that concern sanctions taken at the end of the previous month. Throughout the thesis, I will posit that the application_date variable ("the date from which the restriction applies" in the documentation) corresponds to the day when the moderation action was performed, because it seems absurd (and useless) to moderate for a later time, especially knowing that a vast part of this task is done via automated means. When investigating the TDB, a particular attention was given to four elements:

> (E1) Consistency: How close are self-reported moderation patterns across Meta platforms?
>
> (E2) Evolution: Have (and if so, how have) Meta's self-reported content moderation changed through time?
>
> (E3) Insightfulness: Do major events (a) and major policy changes (b) transcribe onto moderation patterns?
>
> (E4) Justifications: How does Meta justify moderation actions?

(E1) and (E2) investigate whether the database increases the amount of information disclosed, while (E3) and (E4) test elements contributing to meaningful transparency.

Of course, there are a number of limitations to this study. The most obvious limitation to the collected data is that it is almost exclusively produced by Meta itself to report on its own performance, i.e. the data is almost exclusively self-reported. While this is often a flaw to address a research question, it is almost the object of analysis of the present study, which precisely seeks to assess to what extent the data put forward by Meta to promote transparency, voluntarily and as part of the DSA, is reliable and reflects its actual degree of transparency. Still, we remain limited by what is made publicly available by Meta and the European Commission. A second limitation pertains to the difficulty of assessing the "causality" of the DSA over the provisions that are analysed. To begin with, it is extremely complicated to get detailed information about the past practices of Meta, especially for provisions that are embedded onto the platform (notices, etc.) or that do not follow a specific formatting (e.g. AI transparency) because Meta simply replaced this information with its new way of doing. It is even more difficult to access when the goal is not only to know what Meta was doing, but also what users were seeing specifically. The WayBack Machine and newspapers dedicated to social media were used to get the closest possible to having this information, but it was not always possible to access it. The other issue on this matter arises when we observe increases in transparency between December 2020 (date of the DSA proposal) and August 2023 (date of DSA entry into force for VLOPs) that are not particularly framed as responding to the DSA. These increases could well be due to long-planned voluntary measures, but it does not mean that the DSA had no direct or indirect influence over their release either. A final limitation relates to the small number of in-app screenshots that were collected. This may have an impact over the representativeness of the explanations that were exposed and hence prevents us from drawing definite conclusions on user-facing notices.

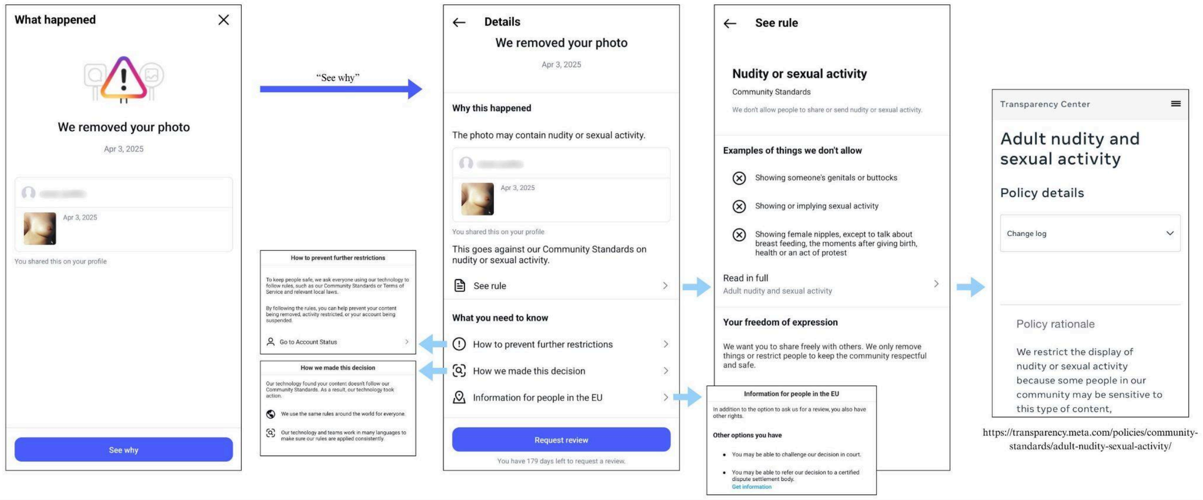
6. Understanding moderation sanctions at the individual level

For transparency to be meaningful at the individual level, users should be able to understand why, whether and how they were sanctioned (cf. section 2.2.). In this section, we make a first step towards bridging methodological challenges (unavailability of user-facing notices, actions, appeals for researchers) to assess to what extent the DSA secures these rights in practice for Meta users. We find that explanations of "whether" and "why" are quite robust, but explanations of "how" have difficulty including processes and hence struggle to get past mere information disclosure.

6.1. Procedural rights: explaining "whether" and "why"

The DSA provides meaningful transparency on *whether* and *why* content has been actioned. By doing so, it entrenches numerous procedural rights that are close to what digital constitutionalism recommends (Palumbo, 2024; Frosio, 2023). Most of them were existing voluntary practices at Meta. The DSA's main contribution at this level is therefore to make these rights mandatory, systematic and, as far as one can tell, equally applied.

First, as regards to understanding *whether* a sanction was taken, the DSA requires platforms to systematically notify users when their content or account has been actioned, regarding visibility, monetary payments, provision of service, or total termination/suspension (Article 17(1) DSA), but also when they have reported someone else's content (Article 16 DSA). Content takedown notices have been applied by Facebook at least since 2016 (Han, 2016), but according to Facebook's and Instagram's own policies they became systematic only between 2018 and 2019 (Gebhart, 2018; Gebhart, 2019). In practice, when Instagram took down the picture posted for this study, it did provide a notice immediately. For the two notices submitted for incompatibility, the delay to receive a notice was six days. Notices are sent in the notifications section and can be permanently consulted in the “Help” section of the settings on Instagram since February 2020 (Hutchinson, 2020), though online research is required to be aware of this feature. Hence, while the DSA does not bring novelty to existing takedown notices, it does make them a legal obligation, which, we showed, increases the value of transparency. Where the DSA innovates is in including demotions and shadow banning in its scope, i.e. sanctions that lower or remove a post’s ability to be visible to other users without informing their owner of the said sanction. Due to its covert nature, the widespread existence of shadow-banning has not been established, but empirical evidence and experiences have shown that it was a likely practice (Le Merrer et al., 2021). If such notices are effectively implemented by Meta, which we were unable to test, the DSA would substantially increase transparency by fostering a better understanding of the platform’s actions (Leerssen, 2023).



What happened

We removed your photo

Apr 3, 2025

Apr 3, 2025

You shared this on your profile

See why

How to prevent further restrictions

To keep people safe, we ask everyone using our technology to follow rules, such as our Community Standards or Terms of Service and relevant local laws.

By following the rules, you can help prevent your content being removed, activity restricted, or your account being suspended.

Go to Account Status

How we made this decision

Our technology found your content doesn't follow our Community Standards. As a result, our technology took action.

We use the same rules around the world for everyone.

Our technology and teams work in many languages to make sure our rules are applied consistently.

Details

We removed your photo

Apr 3, 2025

Why this happened

The photo may contain nudity or sexual activity.

Apr 3, 2025

You shared this on your profile

This goes against our Community Standards on nudity or sexual activity.

See rule

What you need to know

How to prevent further restrictions

How we made this decision

Information for people in the EU

Request review

You have 179 days left to request a review.

See rule

Nudity or sexual activity

Community Standards

We don't allow people to share or send nudity or sexual activity.

Examples of things we don't allow

- ⊗ Showing someone's genitals or buttocks
- ⊗ Showing or implying sexual activity
- ⊗ Showing female nipples, except to talk about breast feeding, the moments after giving birth, health or an act of protest

Read in full

Adult nudity and sexual activity

Your freedom of expression

We want you to share freely with others. We only remove things or restrict people to keep the community respectful and safe.

Information for people in the EU

In addition to the option to ask us for a review, you also have other rights.

Other options you have

- You may be able to challenge our decision in court.
- You may be able to refer our decision to a certified dispute settlement body.

Get information

Transparency Center

Adult nudity and sexual activity

Policy details

Change log

Policy rationale

We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content,

https://transparency.meta.com/policies/community-standards/adult-nudity-sexual-activity/

Figure 5: Path of a content takedown notice on Instagram

Regarding *why* a sanction was taken, Article 17(3) DSA mandates platforms to provide, along with the notice, a statement of reasons describing at least the nature of the sanction (a), the circumstances of the decision (b), the contractual or legal article which the content violates (d-e), and “where relevant” its duration (a). From the screenshots collected, we can see that the sanction is clearly stated (e.g. “We removed your photo”; “You can’t start or join live videos”), as well as the period for which it applies (“Ends on Apr 18, 2025” for the latter restriction). Likewise, the two Article 20 notices are explicit (“We didn’t remove [account_name]’s comment”, “We reviewed [account_name]’s comment and found that it doesn’t go against our Community Standards”). Focusing on the post removal (see Figure 5), we observe that the explanation is present, though thin (“This photo may contain nudity or sexual activity”), but the reference to the contractual ground is explicit: the user has the possibility to click on “See rule” to get a direct link towards the relevant section of the Community Standards in local language. Since this case is easy to understand (female nipples are explicitly forbidden by the platform’s ToS in most cases), especially since it was uploaded with the specific purpose of being taken down, this explanation is sufficient. However, in more ambiguous cases, or for more ambiguous categories than nudity (e.g. spam), the simple assertion of the violated category, though necessary, may not be enough to fully grasp the rationale behind the takedown. Still, this is an improvement compared to 2020 practices as described by Vaccaro et al. (2020), though it cannot be directly attributed to the DSA. At the time, social media platforms were giving minimal explanations of removals, reportedly to prevent malicious actors from "gaming the system" (Vaccaro et al., 2020).

The second component to understand *why* sanctions are taken is the ability for users to know the rules that they are supposed to follow (Article 14(1) DSA, Article 23(1-2) DSA). For a long time, Facebook's rules consisted in short general descriptions of violating motives (Facebook, 2011 in WebArchive). This was later expanded to more explicit yet still widely vague sentences (Facebook, 2014 in WebArchive), and it is only in April 2018 that Facebook released the substance of its Community Standards (Bickert, 2018). Today they are available on the Transparency Center in a user-friendly format, with a very interesting voluntary tool, the Change Log, that displays for each policy the incremental changes that were done since 2018. Once again, the DSA guarantees that this information is available, but it has not significantly impacted its substance (Facebook, 2020 in WebArchive), though the Change Log and more detailed information were introduced after the DSA's announcement.

Finally, a crucial element to guarantee a transparent process and procedural rights, which should necessarily come after understanding "whether" and "why", is the possibility of appeal. The DSA mandates it for users who got sanctioned but also for users who reported content, for a period of at least six months (Article 20 DSA). Instagram's notices present a highly visible button on the bottom to easily appeal the decision, though it is only available for a little less than six months (180 days) and users cannot add any free-text information. Additional information on redress (Article 21 DSA) are also available in a clickable window. Though this seems to be a deeply-entrenched process, it was not not the case when the DSA was introduced. Appealing accounts or pages has been possible for a long time on Instagram's and Facebook's Help Centers, but appealing individual posts was only introduced in 2018 on Facebook, starting with only a few categories (Bell, 2018), and in mid 2019 on Instagram (Hutchinson, 2019), with in-app appeals available as late as 2020 (Hutchinson, 2020). As they are, current mandatory appeal processes are therefore useful to guarantee a transparent process, but they may lack meaningfulness if no justification can be added compared to the first occurrence. Plus, making this process systematic also changes the nature of the process itself: while Facebook VP explained in 2018 that appeals would be "always [reviewed] by a person" (Bickert, 2018), it can obviously not be the case anymore, though it is not to be blamed on the DSA.

6.2. Enforcement procedures: explaining "how"

While the ex ante knowledge of rules (Community Standards) and the user's ex post rights (right to information, right to contest) are quite thoroughly guaranteed in the DSA, there is less clarity on *how* sanctions are enforced, which is yet the third component of understanding sanctions at individual level. Users get some information directly in their notice, but this transparency is more about disclosure than meaningfulness since few processes are explained, including outside the app. Yet as exposed in section 2.2., information on the training and guidelines given to human moderators, information on the creation, operational functioning, and auditability of automatic and AI systems, as well as information pertaining to how these two components of moderation can fail, should be available for transparency to be meaningful at this level.

To begin with, Article 17 notices are supposed to include, "where relevant", information on the use of automated means (c). In practice, by clicking on the "How we made this decision" button, users are informed that the detection and decision were performed through automated means, but the explanation does not go much further. In the flagging notice, this information is even not included: it only contains the impersonal "Our team reviewed the comment". Though it is one of the 2018 Santa Clara Principles, to which Meta adheres, the feature was first tested only in mid-2021, following, Meta says (Broxmeyer, 2021), advice from the Oversight Board (not the DSA).

Outside individual notices, the DSA requires that intermediary services providers publish the measures taken to provide training and assistance to moderators in their transparency reports (Article 15(1)(c) DSA). VLOPs must add the human resources dedicated to content moderation in the EU, broken down by official EU language (Article 42(2)(a) DSA), including their qualifications, linguistic expertise, training and support that they receive (Article 42(2)(b) DSA). For sanctions performed using automated means, intermediary service providers should include the share of article 16 notices processed using automation (Article 15(1)(b) DSA), as well as a qualitative description, a specification of the precise purposes for which automated means are used, their indicators of accuracy and possible rate of error, and any safeguards applied (Article 15(1)(e) DSA). VLOPs are expected to break down accuracy indicators by each official EU language (Article 42(2)(c) DSA).

On paper, these provisions check many of the aforementioned boxes, though we can identify some shortcomings in the legislative text itself. First, while information about the size and working conditions of the moderation staff matters, it says very little on how these moderators enforce the rules. Yet Roberts (2018), De Gregorio (2020) and Vaccaro et al. (2020), among others, have underlined that there is a gap between theoretical rules and how they are enforced. In other words, this information *describes* what available resources the company has to moderate (how many, how trained, how linguistically qualified) and what it implements to address the difficulties of the job (support), but it does not *explain* the guidelines and process through which these people moderate, nor does it outline what is expected from them. This leads to the second limitation: the DSA accounts only for algorithmic failure, not human failure. This does not necessarily mean publishing an accuracy rate for human moderation, though this is what the group of scholars hired by Facebook to independently assess its practices mandated (Bradford et al., 2019), but at least to explain how failure and the unexpected are dealt with. Finally, it is worth underlining that the focus on the description of human and automated "workforces" silences the way that these two means of moderation work together. Once again, this could be addressed with greater emphasis on explanation.

Still, the DSA shows potential for increasing transparency especially given that Meta had relatively few voluntary initiatives on this topic before. When Facebook published its Community Standards in April 2018, these were presented as "the internal guidelines" that content reviewers were following (Bickert, 2018). Surely, they were greatly operationalising Facebook's public rules which had previously been overly broad: "We may also remove support for violent organizations" (Facebook, 2011 in WebArchive), "You may not credibly threaten others, or organize acts of real-world violence" (Facebook, 2014 in WebArchive). Apart from this, Facebook "attempt[ed] to dispel the mystery" (Silver, 2018) on how content moderation is enforced in a blog post article in Summer 2018. It revealed the number of people in its safety and security team (20000, with 7500 content reviewers, increased to 30000 and 15000 reviewers at the end of 2018), of languages (50) and of locations (20), but also its hiring criteria (language proficiency, cultural competency, resiliency), its three-step training process and its mental health support resources (Silver, 2018). This information has not been centralised on the Transparency Center, though specific sections like "How We Create and Use Market-Specific Slur Lists" include a human component. As for automated detections and decisions, there is also relatively little detail compared to Meta's other models.

However, if we analyse Meta's enforcement of the DSA on this topic, we find that the information disclosed in its DSA transparency reports is substantially the same as the one disclosed in 2018, with a greater emphasis on the EU and less detailed procedures. Both texts discuss the same topics (quantitative presentation, training, support, expertise, diversity, audits). However, when the transparency report uses general expressions like "in-depth" and "extensive" training focused on "operational proficiency", the blog post generally displays more specific information, such as a training of "a minimum of 80 hours with a live instructor". The report only adds the existence of two tools available to moderators, the number of global language agnostic reviewers and the number of moderators per EU language. It notably fails to expose the process undergone by a piece of content after it is flagged, even though the blog post addresses this point, but it is also not required by the DSA. Plus, outside commas and figures, the explanations provided in sections 6 and 7 are the same throughout the three transparency reports, except for one sentence deleted in the last iteration which touched upon the number of global content reviewers and their different contractual natures. Finally, even though it is an improvement compared to a 2018 blog post, it can be contemplated whether displaying this information in a legislative plain-text PDF is meaningful enough if we consider that users should be the target audience. Thus, the DSA increases transparency by making the display of this information mandatory but the extent to which users are able to understand how sanctions are enforced stays extremely limited.

7. Understanding moderation sanctions at the aggregate level

The problem with focusing on individual-level understandability to assess meaningful transparency lies in the absence of oversight onto the platform's moderation *patterns*, that is, onto what moderation looks like at scale. As we've discussed in section 1., transparency's benefits not only lie in providing users with the means to understand the space they are in, but also in ensuring that platforms do not abuse their undue asymmetrical power. Fixating on particular cases can obfuscate the larger moderation trends that result from executive decisions and prevent the exposure of commendable or harmful enforcement. Therefore, content moderation transparency should also allow some degree of understanding of moderation at the aggregate level, at least to researchers and regulators, but more broadly to anyone interested. "Aggregate level" is understood as the sum of individual decisions, both in substance (the outcome that they reach) and process (how they are reached overall).

For that purpose, the DSA contains two tools: the production of transparency reports (Articles 15, 24, 42 DSA) and the DSA Transparency Database (TDB) (Article 24(5) DSA). Transparency reports require what has now become an industry standard: the provision of periodic consolidated statistics on the sanctions performed over a given time period. On the contrary, the TDB is an unprecedented mechanism that has no self-governance equivalent. The TDB is a public EU-managed database, paid for by a supervisory fee imposed on VLOPs and VLOSEs (Article 43(2) DSA), in which online platforms are required to log all the (anonymised) statements of reasons that they provide to individuals when they have performed a moderation action. This innovative provision is the literal translation of being able to oversee the sum of individual sanctions enforced by platforms. The database contains 37 variables, about half of which are optional, that range from the nature of the sanction to explanations of its circumstances. A more substantial overview of the database can be found in Annex I. In this section, we dig first into the TDB, then into transparency reports, and finally we try to cross-check the two. We find that at the aggregate level, transparency tends to stop at information disclosure because of the self-declaratory and public nature of the provisions, inherently and in practice. However, comparison could be the key to overcome these limitations and foster meaningful transparency at this level.

7.1. The Transparency Database: a unique but incomplete overview of sanctions

Assuming that the information provided by platforms is reliable, the TDB undeniably offers interesting insights into the overall moderation of platforms. Thanks to this tool, it is possible to see what moderation looks like on a given platform, how it evolves through time (E1), but also what the similarities and dissimilarities across platforms are (E2). Since the data is unexploited and the tool unprecedented, we dedicate the first sub-part to presenting the insights revealed by the TDB. Then, we proceed by showing that inherent limitations as well as implementation shortcomings reduce the degree of transparency that a tool like the TDB could theoretically provide.

7.1.1. Exploring the Database: four lessons about Meta's moderation

The TDB teaches us four main things about Meta's moderation that were previously unknown, though some have been the subject of heated discussions.

First, it provides information about the swiftness at which Meta moderates. Most sanctions are taken within 24 hours after the piece of content has started to be hosted, but the degree to which this is true depends on the platform. Facebook has the harshest moderation: 85% of registered sanctions were taken on the same day as the piece's publication, which implies within the first 24 hours, while the next 24 hours drop to hosting only 2,3% of sanctions, and the next 1%. On Instagram, moderation is slower and more spread over the first few days: 60,8% of sanctions are taken within the first 24 hours, 8,7% in the following 24 hours, and 5,7% in the next. Threads follows more or less the same pattern. WhatsApp Channels does not work the same way, as we will witness several times in this analysis: only 31,75% of sanctions take place in the first 72h. Strikingly, the intelligence never ceases since some of them are actioned up to 21 years after publication. About 5% of sanctions were imposed more than 500 days after the piece's publication on Instagram, 2,09% on Facebook. The longest time frames recorded in this sample were 505 days for WhatsApp Channels (1,38 years), 5220 days for Threads (14,29 years), which is surprising given that it is said to have first launched in 2019, 5262 days for Instagram (14,4 years) and 7652 days for Facebook (20,95 years). These figures go back to Instagram's first three days and Facebook's first fifty days at Harvard, since the services were respectively created on October 6, 2010 and February 4, 2004, i.e. 5265 and 7701 days before the last date of this sample (March 06, 2025). This raises unanswered questions as to how content is flagged for review.

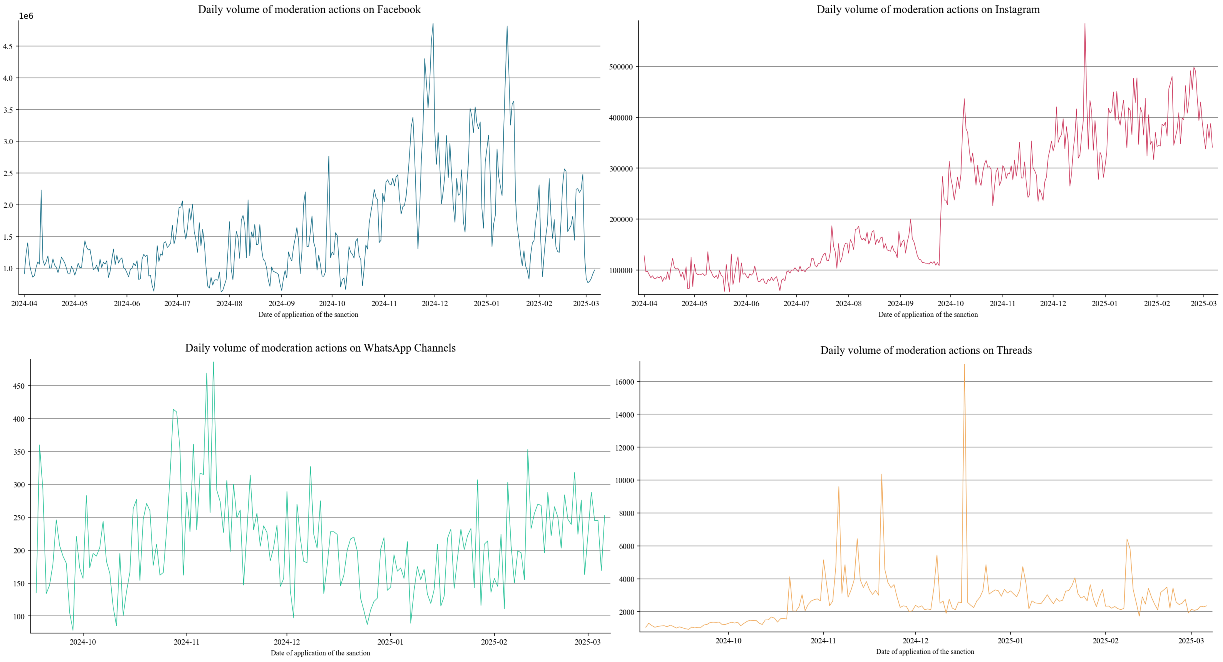


Figure 6: Daily volume of moderation actions across Meta platforms

Second, the TDB shows that the volume of daily sanctions is not constant but rather relatively volatile, and differs between platforms that are yet relatively the same size (see Figure 6). Facebook, with its 260,6 million monthly active EU users (Meta, 2024d), reports six to ten times more moderated items than Instagram and its 269,1 million monthly active EU users (Meta, 2024e). Facebook generally oscillates between a bit less than 1 million and 2,5 million sanctions per day, but can go up to almost five million like it did twice, on November 30, 2024 and January 13, 2025, while Instagram never goes past 600 thousand sanctions per day. Yet Instagram's actioned volume is growing, contrary to Facebook which grew but suddenly came back to previous levels. Threads reports a lot less moderation actions but it also probably has less monthly active EU users (it had 320M users worldwide in January 2025 (Hutchinson, 2025)). As for WhatsApp Channels, it also has less monthly active users than Facebook and Instagram, but its lower volume of sanctions can above all be explained by its different network model. Channels are different from feeds and stories. Further, WhatsApp fosters an image of confidentiality and minimal interference, even if it theoretically does not apply to its non-encrypted channels.

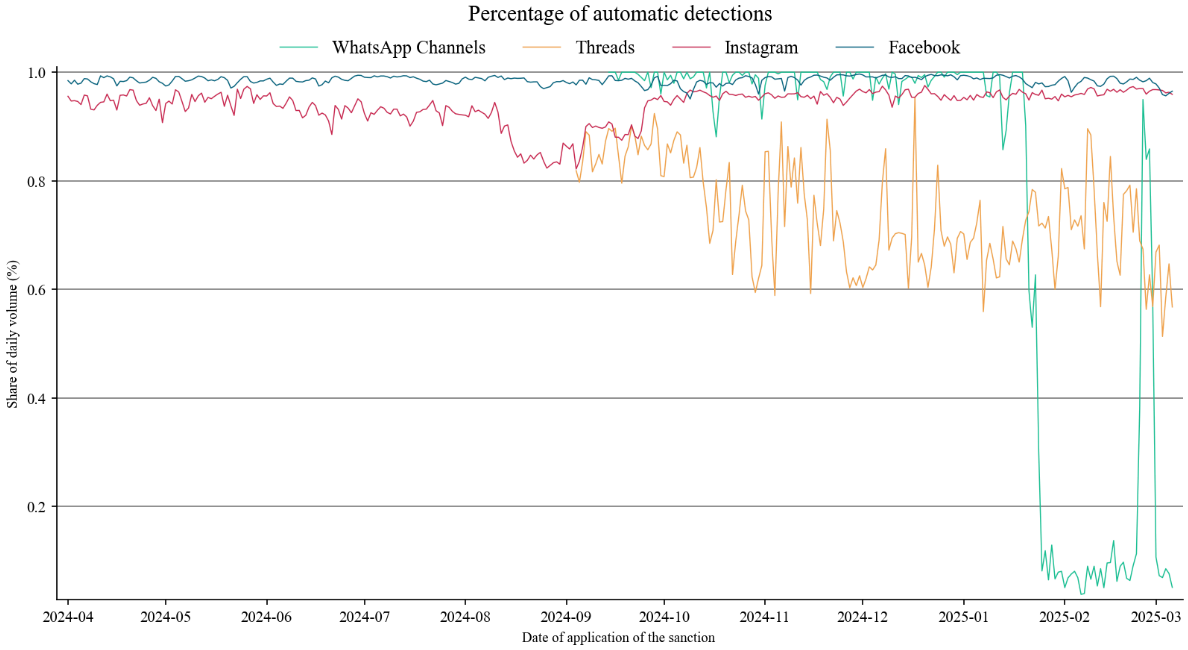


Figure 7: Automated detection rates across Meta platforms

Third, the TDB reveals that moderation is indeed highly automated (see Figure 7), especially on Facebook, where on average 98,5% of detections are automated. On Instagram, it is 93,9% and on Threads, 73,9%. This average hides differences across categories: categories like data protection and privacy violations (+99% automated) and scope of platform services (+99% automated on Facebook, +96% on Instagram) are more automated than categories like intellectual property infringements (55% automated on Facebook, 23% Instagram), non-consensual behaviour (about 80% automated) and negative effects on civic discourse or elections (89% Facebook, 27% Instagram). Interestingly, these three platforms declare that their automated detections perfectly correlate with their partially-automated decisions, and reciprocally with non-automation. This would imply that there are two separate channels of moderation working in parallel, one human and one automated, but we will see that it seems unlikely. WhatsApp Channels is less automated (77,6% of sanctions are), which is consistent with the observation that it is less swift at removing content.

A final element that was known but that the TDB confirms is that moderation is almost entirely voluntary (rather than notified by a user, entity or trusted flagger) and realised upon incompatibility (rather than illegality). The figures indicate that at least 98% of sanctions are taken voluntarily on Instagram and Facebook, which drops to a minimum of 87% for

WhatsApp Channels. As for incompatibility, at least 99,75% of sanctions are taken on this ground across all platforms. Accordingly, given that territorially constrained decisions are all based on illegal content, almost every decision logged applies to the whole EEA. There are only 8000 sanctions on Facebook and 2400 on Instagram over the eleven-month period that have a different territorial scope, almost half of them applying solely to Germany, which is known to have stricter national speech rules.

7.1.2. A flawed and obscure glimpse of reality

Despite these apparent information gains, the extent to which the TDB can help *understand* what happens within Meta, and not merely *see through* its practices, remains limited. The first and most obvious impediment lies in the database's trustworthiness, which flows from its self-reported character. It is unlikely that the information on the TDB is entirely fabricated, though it is theoretically possible. What is likelier is that the information presented is taken out of context, corresponds to the most positive framing possible of the company within this closed environment, or is only partial, and hence shows a skewed version of reality that we risk taking as an objective picture. Three elements of Meta's SoR point to these three shortcomings. Starting with the lack of context, we find that even the simplest figure, volume, could well be a skewed metric. On its website, Meta explains that it counts Instagram and Facebook takedowns differently in its voluntary transparency reports: deleting a post with three images and a caption on Instagram counts as one deletion, but deleting the same post on Facebook counts as four (Meta, 2023b). If the same rules applied to the TDB, this could explain the difference in volume observed between the two platforms, yet there is no way of knowing. Given the amount of sanctions that deal with accounts on both platforms (see Figure 8), we could hypothesise that such a difference does not originate in organic content takedowns but rather on an exceptional amount of probably fake or violating accounts on Facebook. Whichever is true, this shows that metrics taken out of context can be misleading. The second element is that Facebook, Instagram and Threads report no instance of fully automated decision. They always use the two values "not" or "partially" automated. Of course, we should strongly doubt that this is true: it would be difficult for the 5548 human moderators that are reported to be working on the EU to be participating in the average 96,2% of decisions that are taken "partially" automatically while actioning 3,8% of the sanctions themselves (Instagram and Facebook combined). Plus, the notice received on Instagram, analysed in section 6, says itself that the decision was automated ("our technology took action"), which is consistent with the fact that the picture was taken down in two minutes. It is also corroborated by Meta's own website "Our technology automates decisions for certain areas" (Meta, 2025b). However, this label is likely not a blunt lie either: this technology, which is machine learning algorithms for the most part (Meta, 2024g), improves as human moderators make their decision, implying a co-construction of content moderation decisions between humans and technology. This narrative can be found on Meta's Transparency Center (Meta, 2024g) and in its transparency reports (Meta, 2024e, p.19). Given the fierce criticism that relying on machine learning for content moderation receives from scholars (Gorwa et al., 2020), Meta's interest probably lies in emphasising that humans are still in the loop. One thing that should be noted, though, is that WhatsApp Channels surprisingly reports no "partially" but instead "fully" automated items. However, this could once again be explained by the fact that the network acts only on a small volume of content and does not intervene much on speech but rather on explicit images, where the use of automation is less contested. Hence, this label is only presenting reality in a way that is more advantageous to the company. Another example that shows that the TDB might not fully reflect reality is the total absence of end dates in the TDB, which means that all the sanctions that are logged reflect permanent sanctions. Yet we know that Meta imposes temporary sanctions, not least because one of the screenshots collected for section 6 received this kind of sanction. Actually, Meta calls these "strikes" and counts them as punitive warnings before taking a permanent sanction in case of persistent recidivism (Meta, 2024i). It is questionable whether these sanctions should be part of the TDB, but they could be considered partial suspensions of the provision of the service, in which case they would be covered by Article 17(1)(c) DSA. In any case, it is important to have in mind that in the case of Meta, the TDB does not show the full picture of moderation but rather the part of moderated items that were subject to permanent sanctions.

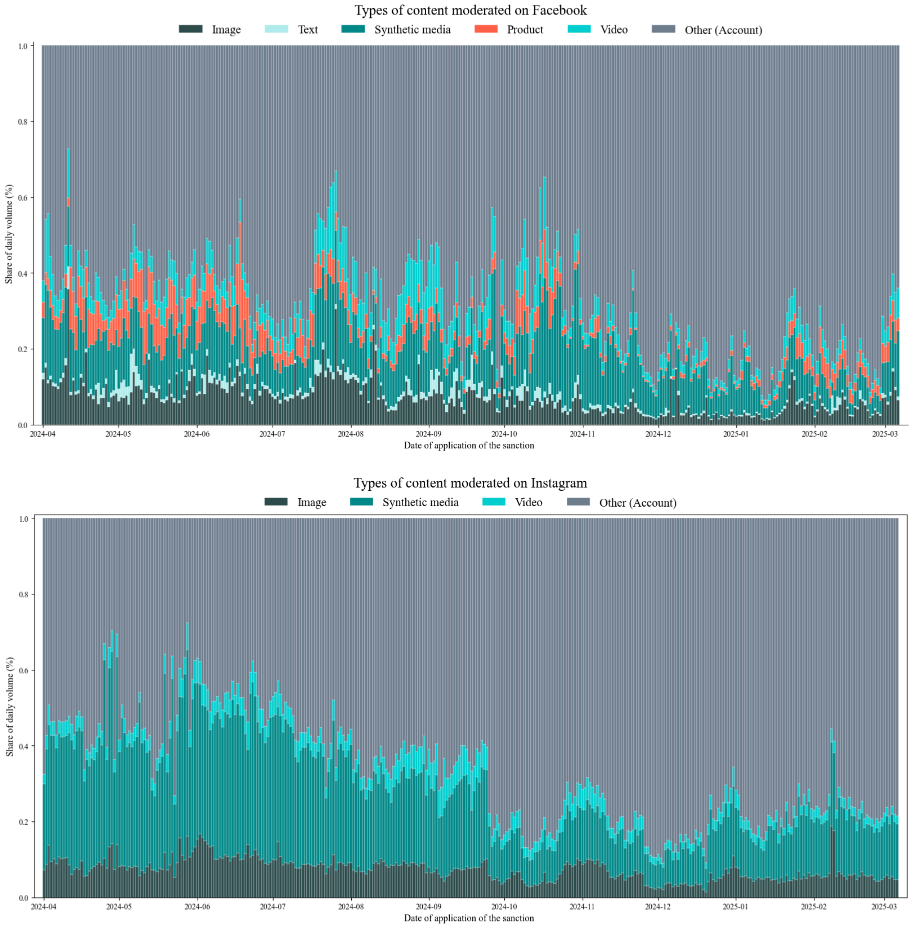


Figure 8: Daily moderation actions on Facebook and Instagram split by content type

Second, there are further obstacles to understanding moderation practices that are due to how Meta fills the TDB. One of the core variables of the TDB is the mandatory "category" variable, which contains the categorial ground on which the piece was actioned. However, what we observe is that the most frequently logged category is "Scope of platform service", which is a large and eclectic category that refers to age-specific restrictions, geographical requirements, goods/services not permitted to be offered on the platform, language requirements and nudity (European Commission, n.d.a). It is the case of Meta (see Figure 9) but also of other platforms (Drolsbach & Pröllochs, 2023; Trujillo et al., 2024), to the extent that it is the first category in the TDB overall (European Commission, n.d.b). As Trujillo et al. (2024) have already remarked, this category is particularly uninformative because of how varied it is but also because it overlaps with other categories (i.e. Pornography or sexualized content, Illegal or harmful speech). Furthermore, analysing Meta's logs through time demonstrates that Meta reports an increasing proportion of its sanctions under this category. A part of it results from a higher volume of sanctions that fall into this category, but even when volume drops to previous levels, the prevalence of this category does not. Therefore it is unclear why exactly the content was actioned, and why this category was chosen over another one. Another obscure filling of the TDB pertains to explanations (E4). First, every optional opportunity to provide explanations or context is empty. Some would have been admittedly costly or heavy to provide, such as providing a URL to the violating section or the specific law that an illegal content violates. However others are information that the platform has but does not give, especially the piece of content's language and the account type (personal or business). Second, the explanations that the company does log, because they are mandatory, have no informational value. Either they are redundant with other columns, or they are generic. For example, the two content explanation variables include entries of the following type: "This was DECISION_VISIBILITY_CONTENT_DISABLED because it violated DECISION_GROUND_ILLEGAL_CONTENT.", whose capitalised expressions are the exact values logged in previous columns. Essentially, Meta explains why the content was considered illegal by saying that the content was disabled because it was considered illegal. A different example, but which leads to the same outcome, is the two variables that expose the grounds on which the decision was taken (i.e. what the piece of content violates). On Instagram, every entry of these variables contains "This was a violation of "How You Can't Use Instagram section" [for illegal content, "of our terms of use" is added]". The said section explains what people cannot do on Instagram, including violating the Community Standards, violating the law or interfering with the provision of the service. Hence, Instagram essentially explains that the content was incompatible with its terms of use because it violated the part of its terms of use that imposes rules on content (its Community Standards). This does not bring any information, while Instagram presumably most of the time has the information since it provides violating users with a link to the relevant chapter of its Community Standards in its user-facing removal notice, like section 6 shows. The same rationale applies to other Meta products.

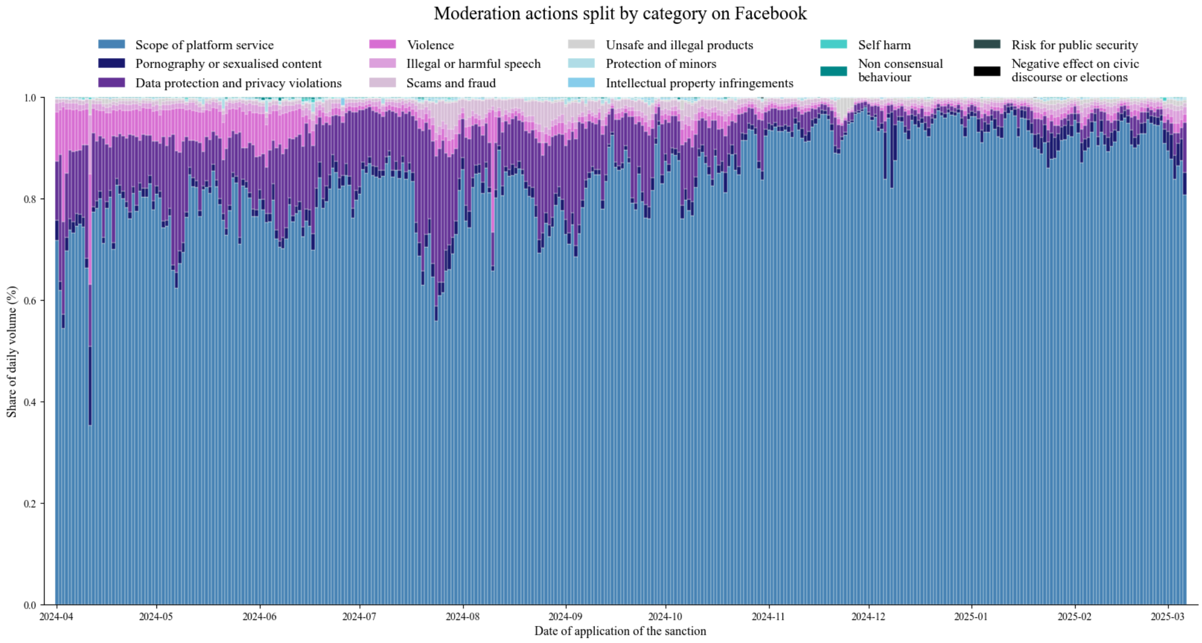


Figure 9: Daily moderation actions on Facebook split by category

Finally, despite the valuable information that it provides, the TDB can be obscure when it shows tendencies that are unexplainable. Some patterns were observed that find no explanation: the sudden appearance of monetary restrictions on October 06, 2024 on Facebook and Instagram, the increase in the number of accounts actioned that violate the scope of platform services in November-December 2024 on Facebook, and above all the sudden and permanent shift on September 26, 2024 in the number of content actioned on Instagram. No information was found on the internet to justify such shifts. However, this flaw might not be the most severe one if we consider that the TDB, as a publicly accessible tool, should "serve as a first-warning system for more targeted efforts" (Leerssen, 2020, p.39), provided that these more targeted efforts are possible (which will notably depend on the not yet implemented Article 40 DSA that grants vetted researchers with access to platform data). Conversely, major real-world events that could have been expected to impact moderation patterns are not observable in the TDB (E3). A major shrinkage seemed to be observable on January 7, 2025, the day Mark Zuckerberg announced his new policies, but after investigation it was only due to a postponement in the logging of January 7's statements onto the TDB, not a change of moderation practices.

In a nutshell, the TDB does increase transparency on Meta's practices because it subjects the company to an unprecedentedly granular and almost real-time oversight of its overall sanctions, that anyone interested in exploring the available machine-readable data can undertake. However, there are two major caveats: first, the case of Meta suggests that some platforms could be complying to the minimum extent possible, in such a way that the TDB's intentions do not match its implementation; second, the data is not representative of Meta's overall moderation, to an unknown degree that would deserve further investigation. This impedes the database's usefulness, as it threatens its informational value and its reliability.

7.2. Transparency reports: a case of obscure transparency

The second provision in the DSA that is intended to increase understanding of moderation sanctions at aggregate level is transparency reporting. All intermediary service providers are required to publish a transparency report once a year (Article 15 DSA), except micro or small enterprises, that contains specific information outlined in §1. Online platforms (Article 24 DSA) and VLOPs (Article 42 DSA) must each include additional information, and VLOPs must publish a report more frequently (at least every six months). Transparency reports contain figures such as those recommended by the Santa Clara Principles on Transparency and Accountability in Content Moderation (2018), for example the total number of pieces of content actioned, accounts suspended and appeals lodged within the reporting period. They also include textual explanations around processes, but the emphasis is put on quantitative metrics. Even though DSA transparency reports seem to be giving a lot of substantial and varied information throughout their twenty to thirty pages, the empirical analysis of Meta's April-September 2024 report finds that the amount of enlightening information is fierce. Plus, the report faces the same limitations as those outlined in the previous subsection: the information provided does not fully encapsulate reality, though it is tempting to believe so, and is presented and selected in such a way that is beneficial to the company's public image.

First, the extent to which DSA transparency reports increase meaningful transparency is limited. To start with, transparency reports had become an industry standard before the DSA entered into force (Reid and Ringel, 2025), so the provision could be redundant with already existing information. It is the case with Facebook (and later Instagram), who has been issuing Community Standards enforcement reports for eight years. Though Meta publishes new information in its DSA reports, for example the overturn rate split by category, three of its five voluntary metrics are taken over (appeals, restored content, content actioned), reducing the extent to which the DSA can "increase" transparency. However, what is most important is that tech transparency reports *in general*, though praised by some scholars in theory, are found to poorly improve transparency in practice (Reid et al., 2024b), at least when they are voluntary. They provide only fractional transparency, that is, by disclosing particular pieces of information, platforms purposefully divert attention from others (Reid et al., 2024b). Transparency reports then become visibility management mechanisms "by which companies seek to maintain legitimacy" (Reid et al., 2024b, p.96) and risk becoming transparency theatres. For example, as discussed in section 6.2., the free text sections of Meta's DSA transparency report are for the most part very generic and vary little from one report to another, both across time and products, yet they are marketed as contributing to transparency. Plus, the emphasis on aggregate figures is misleading as these say nothing in themselves, but are informative only when put in perspective with one another. By removing relevant information, or simply by not providing any percentage, Meta makes this harder.

Furthermore, consistent with the idea of transparency reports as visibility management tools, several aspects of Meta's DSA transparency report limit the extent to which external oversight is possible. First, the report is presented as a PDF and is thus not "machine-readable" as mandated by Article 15(1) DSA. Though Urman & Makhortykh (2023) have underlined that the term lacks clarity, there is a general understanding in computer science that “machine readable” refers to “a data format that can be automatically read and processed by a computer” (Machine readable, n.d.), which is not the case of a PDF. This significantly hampers the ability to analyse and compare the report’s metrics scientifically. Second, also on the format, Meta presents multiple times tables displaying detailed data on only 9 violations out of its 24 Community Standards rules, without commenting on the missing categories nor even on the choice of these categories. This results in tables revealing only a part of the moderation, which opens up legitimate questions on the reliability of the information provided. For example, Facebook’s Table 15.1.c.(1)’s last row reads “Total (including other violations)” with a figure of 49,4 million removals, yet the sum of the rows above only amount to 13,7 million, which means that almost 36 million removals are hidden from view. The obscured categories mostly pertain to integrity and authenticity (impersonation, misinformation, inauthentic behaviour…) and intellectual property, but there is also human exploitation or fraud. This exemplifies how information can appear transparent while being obscure. Finally, it is also possible to decipher in the choice of certain metrics a desire to retain control over sensitive information and public image. It is the case of the accuracy of automated systems: Meta provides only one metric, the Automation Overturn Rate, to measure the accuracy of its automated systems, for Instagram and Facebook combined. In addition to being a questionable choice to measure accuracy, whose limits are acknowledged by Meta, it is ambiguously defined. The “percentage of content actioned using automated means that are later restored” does not say what automated means are involved, nor do they say the reason why content is restored (did a human restore the piece? Did another automated system double check? Was this check triggered by an appeal? What is the checking system’s own accuracy rate?). This metric is either a mean for several systems or the accuracy rate of only one system while Meta presumably has several dealing with content moderation. Plus, it seems that it could have been possible for Meta to publish two different rates (one for Facebook and one for Instagram), as Meta publishes differentiated figures on content actioned and content restored. These clues suggest that this is a kind of *resistant* transparency (Annany and Crawford, 2018): the company does display information but in a way that minimises scrutiny. This interpretation is corroborated by Meta’s two infringements on this provision: first, it did not publish *any* accuracy metric in its first transparency report, which EY’s independent audit pinned (EY, 2024); second, it now does not break down this rate by EU language as mandated by Article 42(2)(c) DSA.

In a nutshell, the DSA transparency reports have limitations that are mostly inherent but also pertain to the implementation of the provision by Meta, who is led to optimise its reputation. Hence, they are only partial elements of transparency that can easily obscure what is absent from the report. Interestingly, Meta itself emphasises this point when it says that the article 9 and 10 requests required in the report are only a part of the government requests that it receives, yet it does not underline such limitations for its own tables. In this sense, this provision of the DSA seems to increase transparency when defined as information disclosure but not so much when it is defined as understandability of sanctions at aggregate level.

7.3. Comparing the database and the reports: a boost in meaningful transparency

We demonstrated that both transparency reports (TR) and the TDB suffer from their public and self-declaratory nature and have difficulty proving themselves meaningful. What could change this conclusion, though, is the fact that transparency reports and the transparency database overlap on some dimensions. Through this lens, transparency reports could be valuable in that they might indicate areas to investigate in the database, and conversely the database could be used to audit the overly self-declarative transparency reports. This could constitute the necessary examination of corporate social responsibility disclosures that Reid et al. (2024b) prescribe to avoid simply legitimising corporate power. There is a lot of information in the reports that cannot be crossed with the database, either because it is not included in the database's scope (e.g. appeals, government requests) or because of technical inconsistencies. A recurrent challenge is that the reports display ToS categories that do not match with the values of the TDB's "category" variable. Still, it is possible to investigate a few elements such as Article 16 notices, demotions and removals. In order to know if transparency reports can be valuable when considered holistically with the database, we realise these comparisons for Instagram and Facebook (see Figures 10 and 11).

	Type of alleged illegal content	"Total number of [Article 16] notices resulting in content removal for policy violations" (TR)	"Total number of [Article 16] notices resulting in restriction of access to content due to alleged illegality" (TR)	Total number of SoR with "Article 16" as source, "Incompatible content" as ground and "Content removed" as sanction (TDB)	Total number of SoR with "Article 16" as source and "Illegal content" as ground (TDB)
Instagram	Intellectual Property (TR) / Category "Intellectual property infringements" (TDB)	29 657	0	69 887	0
	Privacy (TR) / Category "Data protection and privacy violations" (TDB)	6 078	11	6 530	0
	Total (all categories)	76 084	2 019	79 035	16
Facebook	Intellectual Property (TR) / Category "Intellectual property infringements" (TDB)	45 643	0	98 575	0
	Privacy (TR) / Category "Data protection and privacy violations" (TDB)	8 477	42	8 334	0
	Total (all categories)	98 899	5 172	120 555	6

Figure 10: Comparison of aggregate Article 16 notices across the transparency report (TR) and transparency database (TDB) for Facebook and Instagram between 01/04/2024 and 30/09/2024

There are systematic differences between transparency reports and the database in volume, though some figures are strikingly close, but automation rates are fairly consistent across mediums, if we consider that the TDB's "partial" automation matches with the TR's full automation (Trujillo et al., 2024). Several elements could explain such differences. They could originate in a difference in the definition of sanctions anywhere between content reviewers, the drafters of the reports or the engineers who enabled the TDB for Meta. This seems plausible given that Facebook's volume of "Account terminations" is extremely low in the TDB and high in the TR, while "Account suspensions" are extremely high on the TDB. Likewise, "intellectual property" Article 16 notices could have been counted as belonging to another category in the TR, though it is surprising given that it is the only one of the two categories whose labels match perfectly between the two. Differences could also originate in mismatching time periods: the TR mentions a "reporting period" from April 01 and September 30 which was interpreted on the TDB as an application of the sanction between these dates, but it could be otherwise. Finally, though it would be troublesome, any one (or both) of the two mediums could be incomplete. The fact that automation rates match could suggest that there is only a proportionality difference between the two, and hence perhaps a random screening between TR figures and the TDB, yet the fact that the TDB sometimes reports higher volume than the TR goes against this hypothesis, which leaves us with no definite response.

		Volume (TR)	Automation (TR)	Volume (TDB)	Automation (TDB)
Instagram	Organic content removal measures (TR) / "Content removed" as sanction (TDB)	12 136 947	10 535 490 (86,8%)	6 513 905	5 523 101 (84,79%)
	Organic content demotion measures (TR) / "Content demoted" as sanction (TDB)	2 035 562	2 035 311 (99,99%)	2 303 743	2 303 743 (100%)
	Account termination measures (TR) / "Account terminated" as sanction (TDB)	17 991 379	16 582 852 (92,17%)	12 755 774	11 982 897 (93,94%)
Facebook	Organic content removal measures (TR) / "Content removed" as sanction (TDB)	49 361 368	46 758 906 (94,70%)	59 247 892	56 854 817 (95,96%)
	Organic content demotion measures (TR) / "Content demoted" as sanction (TDB)	27 021 695	27 010 205 (99,96%)	18 286 756	18 286 756 (100%)
	Account termination measures (TR) / "Account terminated" as sanction (TDB)	87 825 945	79 195 203 (90,17%)	1 020 885	972 139 (95,23%)

Figure 11: Comparison of selected moderation actions across the transparency report (TR) and transparency database (TDB) for Facebook and Instagram between 01/04/2024 and 30/09/2024

A final insight from this comparison concerns Article 16 automation rates. The database shows that 18,38% of Facebook Article 16 reviews and 5,59% of Instagram's are reported as being automated. If it is true, this exposes that the phrase "All Article 16 DSA notices are processed using manual review" is yet another way of managing the company's public image by omitting that "instances of duplicate submissions" are treated by automated means.

All this shows that reading transparency reports and the database against other mediums, especially each other, could substantially increase their value in fostering meaningful transparency. For this specific reason, mandatory and comparable transparency reports could constitute real assets for research. At the same time, this increase in meaningful transparency at the aggregate level for now results in shedding doubt on the reliability of both the transparency reports *and* the database. While the value of the two mediums is enhanced by the comparison, it would be *in fine* considerably decreased if they were indeed found to be incomplete.

8. Understanding moderation at the systemic level

The last level at which content moderation can be meaningful is what this thesis labels the systemic level. It differs from the aggregate level in that it takes a broader perspective on content moderation: first, it considers moderation not only as the sum of individual sanctions but as anything that impacts the organisation of content; second, it seeks to understand not only how moderation impacts the rights of users but how it structurally impacts society and communication at large. At this level, we find that the DSA increases transparency on advertising and especially recommender systems, though there is still room for improvement. For more quintessential accounts of Meta's impact on society, the DSA is not yet implemented enough for us to draw definite conclusions, yet an early look at risk assessment reports suggests that these could miss a part of their intended purpose.

8.1. Towards advertising and recommender systems transparency

The DSA restricts the definition of content moderation to actions sanctioning users, but as discussed in previous sections content moderation can also be defined as encompassing anything that impacts the content that users see. This conceptualisation is especially relevant given the increasing role that (machine learning) algorithms have been playing in ordering content over the past years. Social media feeds have changed, from prioritising content originating from friends to prioritising algorithmic recommendations. According to Meta's own reports, 54,4% of US feed content views in Q4 2024 come from accounts, groups or pages that users are following and 33,6% from unconnected sources (Meta, 2024f) while in Q2 2021 90,6% came from these connections and 8% only from unconnected sources (Meta, 2021). Advertisements are not exempt from this trend. Beyond the available micro-targeting strategies of purchased ads, the platform's ad delivery algorithms themselves have strong influence over what advertisements users are exposed to. Notably, Facebook's ad delivery algorithms were found to be skewed in several studies, perpetuating gender biases by showing more financial ads to men (Ali et al., 2023) or less remunerative jobs to women (Imana et al., 2021), more problematic ads to older people (Ali et al., 2023), or political messages predominantly to likely proponents of the said messages (Ali et al., 2021). This has an impact on user experience, especially given that users were seeing a median of 70 advertisements per week on their Facebook feed in 2019, i.e. between 10 and 15% of posts seen (Arrate Galán et al., 2019).

The DSA includes this perspective when it mandates transparency on advertising and on recommender systems, as detailed in recitals 68 and 70. In the directive, all online platforms must show user-facing transparency within their product on the advertising nature of organic and ad content (Article 26 DSA). They must also explain in their terms and conditions the main parameters used in their recommender systems and the reason behind their relative weight, as well as possibilities for users to act on these parameters (Article 27 DSA). In addition to this, VLOPs and VLOSEs must create and maintain a searchable public repository of all the advertisements that are running or have run on their platform at least one year after their last display (Article 39 DSA), with specific information outlined in the second paragraph such as the content of the advertisement, the identity of the purchaser, the beneficiary of the advertisement, the target audience if there is one and the reach of the advertisement. In the case of Meta, we find that the DSA did foster additional transparency on understanding recommendations and advertising but not to its most meaningful degree.

Starting with recommender systems, the literature review has shown that transparency can be improved mainly by giving legible (visually appealing, concise, summarised) and non-technical explanations on how systems work and by explaining the controls available to users. In the case of Meta, there are four places where we can find recommender system transparency on Instagram and Facebook, two of which have constituted Meta's main voluntary approach before the DSA. First, their respective Help Centers have a dedicated section outlining the product's Recommendation guidelines, which explain in plain text what content is not eligible for recommendations on Facebook and Instagram. This approach to algorithmic transparency is negative in the sense that it defines how it *does not* instead of how it *does* recommend content. It has been Meta's main public-facing voluntary approach, since these guidelines were published in August 2020, months before the DSA was introduced. More positive transparency was introduced on June 29, 2023. The causal impact of the DSA on this worldwide release is not obvious since Meta has not framed its announcement in this way (Meta, 2023c), but it is also difficult to ignore that the DSA compliance deadline for VLOPs, including these specific provisions, was less than two months away (August 25, 2023). Plus, Meta brought it up in its blog post on DSA compliance (Clegg, 2023). The announcement in question is the release of a new section on Meta's Transparency Center containing 22 AI transparency pages ("system cards") where the functioning of Meta's AI ranking algorithms (e.g. acting on comments, notifications, stories) are explained for Facebook and Instagram, with later updates adding Threads and some Meta Horizon products. Each card first succinctly explains the steps that the said systems go through to complete their task, then points to available user controls and finally describes a few of the systems' predictions, along with some of their input signals. These cards represent positive progress towards meaningful transparency. Their structure is clear and harmonised, they are linked in the terms and conditions and are available in a variety of languages, not just English. Further, some efforts were done on non-technical user-friendliness: the first two parts are remarkably concise and some of the technical words are defined (e.g. input signals, predictions). The cards constitute an increase in transparency since little positive information was previously available on AI content ranking. However, there are several limitations as to how much they allow us to *understand* how AI algorithms function. First of all, while they do list criteria used for some significant predictions (Article 27(2)(a) DSA), the cards do not explain why these predictions are important, i.e. "the reasons for the relative importance of those parameters" (Article 27(2)(b) DSA). They also do not mention how important the listed criteria are and whether they all are of equal importance for calculating the prediction, only that “signals influencing this prediction include” those listed. In general, there is a lack of clarity as to why these specific pieces of information were selected over others for display, which diminishes their value as regards to transparency. This is exemplified by instances where the cards are redundant or incoherent. For example, Facebook notifications card shows that the same signals are used for “How likely is it that more notification volume will lead to more activity on Facebook app” and “How likely is it that low notification volume will lead to fewer activity on Facebook app” while similar twin predictions don’t show the same input signals (Meta, 2025d). Alternatively, the exact same prediction can be shown several times within the same card (e.g. How likely you are to click a push notification for Facebook notifications, How likely you are to “like” a post for Instagram Explore) without having the same input signals every time (Meta, 2025d; Meta, 2025c). Second, which the DSA does not mandate, additional information could be given on the training of the models, their assessment and their updates to foster thorough AI transparency. However this could be addressed as part of the AI Act, especially if social media recommendation algorithms are labelled “high-risk” systems (Bayer, 2024), which is out of the scope of this thesis. Finally, the fact that detailed theoretical information is available on a dedicated website is valuable but its separation from the app and from the content in question can be a barrier to understanding ranking. It would be interesting to think about the design of an in-app feature that either provides a part of this information itself or links the website in a more obvious way. Actually, Meta has been partially doing it since 2019 (Meta, 2019a), voluntarily, through its “Why Am I Seeing This?” feature that can be accessed by clicking on the three dots on the top right corner of the post. However, while the explanation is quite substantial for content originating from connected sources on Instagram (three to four concrete criteria that the post matches), there is almost nothing for unconnected sources (“based on many things, including your activity on Instagram”). Both link to a (rather short) page of the Help Center dedicated to recommendations (on the model of WhatsApp Channels’ ranking transparency (WhatsApp, n.d.)), which in turn links to the AI cards.

In the case of advertising, the DSA renders mandatory practices that Meta mostly followed already, but it does also trigger the release of additional details in some areas. Meta’s ad transparency is threefold. First, there is a user-facing component that matches the requirements of Article 26 DSA. Ads have been identifiable by the “Sponsored” label under the account name that presents the ad (fits paragraphs (1)a-b) for a long time. Yet scholars have shown that not all labels are conducive to more transparency since users tend to skip a non-prominent label when they scroll, so mandating the sole presence of a label does not substantially increase transparency. The promotion and updates of standards on the matter (Article 44 DSA) could be a lead. Further, explanations on why the ad is presented is available through a similar “Why Am I Seeing This Ad?” label in the three-dots menu. This has been the case since 2014 with the display of targeting criteria related to demographics, interests and website visits (Tsukayama, 2014; Meta, 2019b) but the current version, which adds criteria related to the user’s activity on the app and a direct link to change targeting parameters in the settings (fits paragraph (1)d), might have been influenced by the DSA since it was introduced in February 2023 on Facebook and October 2023 on Instagram (Pavón, 2023), though it is impossible to assert it. This increases user controls, which was found to be highly positive. Finally, ads contain no in-app mention of who has paid for the ad (paragraph (1)c), but it links to the specific ad's page in the Ad Library, where the information is available. It is not possible to know whether this link was introduced specifically for the DSA or not. Second, the DSA is directly responsible for the release of additional information in Meta's already existing Ad Library. Since June 2023, Meta has been asking advertisers to declare who the beneficiary and payer are when they purchase an ad dedicated to an EU or global audience (Meta, n.d.b) and has been publishing this information in the Ad Library since the end of August 2023 (Hutchison, 2023). Additional information is also available for each ad under the "European Union Transparency" label, displaying the demographic targeting criteria, the total reach and the reach split by location, age and gender. Non-EU audiences (except Brazil[1]) never have access to this information, especially given that non-political ads are not released for the vast majority of countries and when they are, transparency is narrowed down to the content, dates and beneficiary of the advertisement. Access to granular (and hopefully systematic) platform data like the Ad Library offers, especially through its APIs, is especially valuable for researchers. Meta's two APIs have been used multiple times to inform the study of political (Minihold and Votta, 2024) and economic actors (Xiao, 2025) but also to expose Meta's own moderation flaws (Bouchaud, 2025; Bouchaud and Liénard, 2024). Remaining concerns pertain to the self-declared nature of the beneficiary-payer labels, which often do not correspond to the "natural or legal person" (Article 26(1)(b-c) DSA) who paid for the ad (Xiao, 2025), but also to the increasing power that Meta's AI algorithms have over the targeting strategies of marketers (Votta et al., 2024).

8.2. Exposing social media's structural impacts: innovative but still limited oversight

Finally, the most high-level conception of meaningful transparency contends that transparency should be able to reveal what the impact of social media platforms on society and individual behaviour is at large. It is both the level that exposes the most the business model and the underlying motivations of platforms, which can strongly misalign with the public image that they nurture, and perhaps also the hardest kind of transparency to produce as no specific metric or widespread definition exists. There are two mechanisms through which the DSA aims for this level of meaningful transparency. First, VLOPs and VLOSEs are required to produce once a year a risk assessment report to "identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service [...] or from the use made of their services" (Article 34(1) DSA). They also have to present appropriate mitigation measures relating to each of the risks identified in this report (Article 35 DSA). Second, it mandates VLOPs and VLOSEs to grant vetted researchers access to non-public internal data for projects assessing these systemic risks and mitigation measures (Article 40(4) DSA). Hence, one mechanism comes from platforms themselves while one is independent. Two years and a half after the adoption of the DSA, these two elements are still ongoing implementation, limiting the extent to which their effectiveness in fostering transparency can be assessed. VLOPs and VLOSEs have published either one or both of the two reports that they have already produced (September 2023- August 2024 and October

[1] Additional advertising transparency was mandated by Brazil's National Consumer Secretariat on July 30, 2024 through Nota Técnica Nº 2/2024/Gab-DPDC/DPDC/SENACON/MJ (Campetti Amaral et al., 2024)

2022- August 2023). However, the Commission and the Board for Digital Services (composed of national data agencies, i.e. Digital Services Coordinators) have not yet published their synthesis (and their appraisal) of these reports (Article 35(2) DSA). As for Article 40, the Commission is still working on implementation, including on the establishment of a DSA data access portal (European Commission, 2024), so that no researcher has had official access to Article 40 data to date. For now, it can be argued that the new emphasis on systemic risks bears true promises of increasing transparency, but the current limited implementation does not (yet) rise up to expectations.

Whether or not risk assessments were already done by Meta, little information of this type was public. Meta has been emphasising security risks stemming specifically from the misuse of its products on its Transparency Center, and has been publishing "adversarial threats" reports focusing mostly on coordinated inauthentic behaviours, but occasionally on other risks such as malware, since 2017 (year of the Russian interference scandal). Other than that, the closest that can be found to a broader reflection on Meta's social impact is Facebook's series of blog posts in 2017-2019, which directly stems from the series of scandals that it was going through at the time. In these articles, Facebook reflects on "the impact of our products on society" by giving the floor to prominent Facebook figures (e.g. Chakrabarti, 2018), journalists (e.g. Van Zuylen-Wood, 2019) and various stakeholders including academics (e.g. Sunstein, 2018). It also committed to releasing the slides of its bimonthly Product Policy Forum Minutes, where top executives discuss updates of the Community Standards (Meta, 2018a). This initiative gave a sneak peek into what the processes and decision making hierarchy on drafting moderation guidelines are (Van Zuylen-Wood, 2019; Meta, 2018b), which is a component of meaningful transparency at systemic level. Yet this information is now outdated and seems to have only been the result of significant societal pressure. Since then, there have been few updates. For example, the release of Minutes was regular over 2019 only and stopped altogether in 2023 (Meta, 2024h). In this context, Articles 34-35 are promising because not only do they force platforms to be transparent about the systemic risks that they contribute to, but they also allow the Commission to pin platforms if it considers that their due diligence is not robust enough (Eder, 2024). As for Article 40, mentioned multiple times in this paper, its potential contribution is colossal in a context of ever-increasing impediments placed by social media on research tools designed to study their operations (Karanicolas, 2021), the latest move to date being the discontinuation of Meta's tool CrowdTangle used to track disinformation. Yet it is also the reason why platforms could oppose the release of such data, hence the effectiveness of this provision exclusively depends on interpretation and implementation (Söderlund et al., 2023).

Despite incomplete implementation, it is possible to have a look at Facebook's and Instagram's public risk assessment reports (2023-2024). To get an idea of the reports, it is enough to browse only one of the two because they are essentially similar (only the calculation of risks and a few sentences differ). This in itself is problematic, as it posits that Facebook and Instagram face the same risks. In essence, Bernard (2024) emphasises that Meta did put a lot of work into the report, but aside from a handful of inherently interesting information, the usefulness of the report mostly lies in better knowing what data to ask for in the future. He notes that Meta focuses *heavily* on mitigation measures in its report, at the expense of risk assessment, and that the risks that are outlined almost entirely stem from users and not from the platform's design and rationale themselves. These are interesting pieces of information, but they are only a part of what was expected from these reports. Similarly to transparency reports, risk assessment reports might actually suffer from a visibility management problem. Coding for these elements reveals that sub-parts, particularly risk and problem areas, usually follow the same structure: they start with a few sentences asserting Meta's value-based commitment to the particular area, then either directly detail what mechanisms Meta has put in place to safeguard this commitment or explain how malign users try to escape Meta's engagement towards the policy, before detailing further how these threats are addressed. Most of the time, though not always, when Facebook touches on its own role, it is directly next to a risk stemming from users. For example, when it essentially says that its systems can spread some undesirable content, it is talking about violating content that users have uploaded on the platform. Likewise, when a platform-induced risk is mentioned, it is often directly preceded or followed by a mitigation measure. This strategy of sandwiching risks, especially platform-induced risks, in-between mitigation measures shows that there is a strong incentive for Meta to polish its image and show its diligence. Yet this can be done at the expense of transparency, since the way Meta envisions risks is not particularly salient in the report. It can also be done at the expense of clarity, since the report is full of redundant sentences that are slightly modified. A telling illustration is the 114 and 101 uses (out of 85 textual pages) of the roots "improv" and "continu", including ten instances of "continued improvement", four "continue to improve", two "continuously improving" and ten "continuously working to improve and enhance our". Of course, this could be only an issue with the public version of the report, whose similarity with the one sent to EU authorities is unknown. Two final caveats that civil society organisations have emphasised is that it is unclear to what extent civil society was involved in Meta's assessment process (Windwehr, 2025), and more generally that reports are uselessly posted long after the risks have passed (Algorithm Watch, 2024). Overall, Bernard (2024) contends that Meta's report will satisfy "operations and systems aficionados" but not those "seeking a more philosophical treatment" of Meta's appraisal of risks. It is now up to the Commission to assess to what extent this is satisfactory regarding DSA obligations. In terms of transparency, though, a lack of stakeholder investment, a late ex post reporting and obfuscated language may not rise up to meaningfulness.

Conclusion and recommendations

In this study, I show that the DSA has increased meaningful transparency at Meta, but the extent to which this is true varies across levels. At the individual level, the DSA increases meaningful transparency by making voluntary procedural notices mandatory, so that users understand whether and why they have been sanctioned; yet it has more trouble fostering meaningful information on how moderation is done. At the aggregate level, the DSA's two self-declaratory tools rise from pure information disclosure to meaningful transparency when they are cross-checked, even if the outcome is for us to doubt their reliability. Finally, at the systemic level, the DSA strengthens advertising and recommender system transparency;

however, the early look given at risk assessments reports suggests that the DSA only partially manages to increase meaningful transparency at this level.

From these results, we observe that as of current implementation, the more a provision leads to publicly exposing the structural power of the platform, the less the DSA manages to increase meaningful transparency. Indeed while in-app user-facing procedural rights are extensive, processual or structural qualitative explanations, research access and aggregate transparency are to this day more fragile. Opposing forces stem from two sources. Some originate in the DSA itself, which contains blind spots and built-in limitations. Others arise from implementation: first, because they are corporations, social media are led to engage in visibility-managing strategies to safeguard their public image, which leads them to produce empty or distorted information; second, we also observe that they sometimes comply to the minimum extent possible, probably because of a lack of desire for scrutiny.

These results must be recontextualised in two respects. First, they are based on Meta's case only. The extent to which the DSA increases transparency not only depends on the company's previous level of transparency but also on the diligence with which it complies with the legislation. While in general Meta did have the most advanced transparency provisions before the DSA was implemented, which implies that there is greater potential for the DSA to increase transparency in all other companies, the degree to which other platforms will comply cannot be inferred from the present study. Therefore, it would prove extremely valuable to reproduce the same study on other social media platforms to nuance or confirm these conclusions. Second, these results stem from an analysis of objective transparency, yet it can be questioned whether the DSA's effectiveness should be assessed through objective or through users' perceived transparency. As part of this reflection, studying user-perceived transparency, or even researcher- and civil society- perceived transparency, would undeniably bring interesting contributions to the examination of the DSA's effective impacts.

From these results, I issue seven recommendations to three actors, the Commission, academia and civil society, to improve the DSA's ability to effectively foster meaningful transparency.

Recommendations to the European Commission:

1) **Mitigate platforms' obscure transparency practices by raising and clarifying minimal expectations**

We demonstrated that platforms could sometimes comply to the minimum extent possible, whether per se or as a result of visibility-management strategies. Against this backdrop, the natural solution would be to increase the minimum standards that are expected from social media platforms. In particular, the analysis pointed at the following points:

- a) Require platforms to explicitly disclose in their transparency reports whether their processes have changed since the last reporting period.

This would prevent the legitimising effect that a three-page description of processes trigger while they are sometimes identical to previous reporting periods. At the same time, it could also help track changes.

- b) Invite platforms to split accuracy rates by systems, or alternatively to explain what systems go under the "automated means accuracy rate" of their transparency reports.

- c) Create a short platform-specific documentation attached to the Transparency Database where each platform could explain its choices in filling the database, and its understanding of the different variables.

This would make explicit the possible differences of definitions of the same concept (e.g. account suspension) across transparency provisions and across platforms. It would also give platforms the opportunity to justify the rationale behind their choices (e.g. Meta's use of "partial", instead of "full", automation).

- d) Require platforms to log statements of reasons under the database's category which is closest to the infringement in question, instead of logging them under "Scope of Platform Service".

- e) Invite platforms to log more meaningful textual explanations on the Transparency Database.

- f) Clarify if strike-like temporary sanctions should be part of the Transparency Database. If so, inquire why Meta has no temporary sanction logged.

The Commission's Implementing Act on the harmonisation of transparency reports from November 2024 (2024/2835) goes in the right direction. It directly addresses point d) by deleting the category "Scope of Platform Service" and asking providers to use the "Other" category only if the infringement is not better described by any of the other categories. It also addresses point c) by requiring rates to be provided "per type of content moderation system". While a) is not included, the new CSV format could allow readers to more easily spot whether text variables are identical from period to period.

2) Encourage and support independent research

- a) Enforce Article 40 DSA.

This provision was raised several times throughout the analysis. It is currently in the process of implementation, but its potential is such that its importance should be underlined once more by including it in the present recommendations.

- b) Require transparency provisions to be released in both PDF and machine-readable formats, like HTML or CSV.

- c) Invite platforms to release user-facing design choices to researchers.

Even though it is available to users in-app, some information that should be available to researchers is not, or only partially, public. This is especially the case of the various notices that the DSA compels platforms to send. Researchers should not have to be “forced” to breach the platform’s terms of service to find out what these notices look like, especially that notices for illegal content remain inaccessible.

The November 2024 Implementing Act addresses point b) for transparency reports. It is a substantial progress, though this recommendation also applies to the risk assessments reports.

3) **Invite platforms to increase in-app references to transparency provisions**

Reciprocally, some information that should be available to users is in effect only readable by specialists. In particular, information on how moderation is enforced, which scholars see as being of relevance to individuals, is mostly absent from the app itself. Without going as far as adding the information itself, an in-app URL linking to the relevant section of the transparency website could make this information more meaningful to users.

Recommendations to Researchers:

4) **Investigate further the difference between transparency reports and the Transparency Database.**

We have demonstrated that transparency reports and the transparency database rise in meaningfulness when they are cross-checked. Researchers should explore this perspective with other platforms to see if differences are systematic, and if not where they occur. In this regard, the November 2024 Implementing Act will be a game-changer, at least on paper, since it will ask platforms to report along the same categories as in the database, and more comprehensively than is currently being done. Such investigations could also potentially inform Article 40 research requests.

5) **Investigate how shadow-banning is taken into account in practice under the DSA**

In this study, we were unable to test one of the core innovations of the DSA, namely the inclusion of visibility sanctions in its scope, such as shadow-banning and demotions. Exploring how it is implemented in practice (user notices, TDB) will be key to assessing the success of the DSA in increasing transparency.

6) **Leverage web archiving as a research material and a research practice for the study of social media platforms**.

This inquiry suggests that web archives have the potential of becoming precious historiographical resources that can be leveraged against the discursive power of social media platforms. In particular, they can be used as historical evidence to contest the idea that transparency practices have always been evident. Further, archiving can also be a useful research practice, to keep up with the fast pace of social media. During this very study, Meta's transparency website (and the transparency database) changed several times, forever updating some paragraphs. Archiving can prevent these losses, and be used as material for later research. Actually, Ben-David (2020) made a similar case for counter-archiving Facebook. Current attempts of rewriting historical narratives in other areas should urge us to consider saving this information.

Recommendation to Civil society actors, including NGOs and journalists:

7) **Engage with platforms' transparency material and relay findings from academia.**

Civil society bears a key role in ensuring that the DSA translates into actual transparency and accountability. Without discrediting dialogue and collaboration in any way, Meta's case shows that platforms are responsive to pressure, given that Meta's transparency increases have almost systematically stemmed from scandals or reputation crises. Hence, by translating and amplifying findings from their own research or from academia towards the relevant audience in an appealing way, they can help reward or sanction the enforcement and behaviour of platforms.

Bibliography

Academic work

Acemoglu, D., Huttenlocher, D., Ozdaglar, A. and Siderius, J. (2024) *'Online business models, digital ads, and user welfare'*, NBER Working Paper Series, Working Paper 33017. Available at: https://www.nber.org/papers/w33017

Ackerman, J.M. and Sandoval-Ballesteros, I.E. (2006) 'The Global Explosion of Freedom of Information Laws', *Administrative Law Review,* 58, pp. 85-130.

Acs, Z.J., Song, A.K., Szerb L., Audretsch, D.B. and Komlósi, É. (2021) 'The evolution of the global digital platform economy: 1971–2021', *Small Business Economics*, 57(4), pp. 1629-1659. Available at: https://doi.org/10.1007/s11187-021-00561-x

Ali, M., Sapiezynski, P., Korolova, A., Mislove, A. and Rieke, A. (2021) 'Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging', in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery (WSDM '21), pp. 13–21. Available at: https://doi.org/10.1145/3437963.3441801.

Ali, M., Goetzen, A., Mislove, A., Redmiles, E. M., & Sapiezynski, P. (2023) 'Problematic advertising and its disparate exposure on Facebook', *Proceedings of the 32nd USENIX Security Symposium,* August 9–11, 2023, Anaheim, CA, USA, pp. 5665-5682. Available at: https://www.usenix.org/conference/usenixsecurity23/presentation/ali

Ameli, N., Drummond, P., Bisaro, A., Grubb, M., & Chenet, H. (2020) 'Climate finance and disclosure for institutional investors: Why transparency is not enough', *Climatic Change*, 160(4), 565-589. https://doi.org/10.1007/s10584-019-02542-2.

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

Ang, P. H., & Haristya, S. (2024) 'The Governance, Legitimacy and Efficacy of Facebook's Oversight Board: A Model for Global Tech Platforms?', *Emerging Media*, 2(2), pp. 169-180. Available at: https://doi.org/10.1177/27523543241266860.

Angus, D., Burgess, J., Carah, N., Hayden, L. and Obeid, A. (2023) 'Exploring Facebook's "Why am I seeing this ad" feature: meaningful transparency or further obfuscation?', *AoIR Selected Papers of Internet Research*. Available at: https://doi.org/10.5210/spir.v2023i0.13389.

Arrate Galán, A., González Cabañas, J., Cuevas, Á., Calderón, M., & Cuevas Rumin, R. (2019) 'Large-scale analysis of user exposure to online advertising on Facebook', *IEEE Access*, 7, pp. 11959-11971. https://doi.org/10.1109/ACCESS.2019.2892237.

Baldwin, C.Y. and Woodard, C.J. (2009) 'The Architecture of Platforms: A Unified View', in A. Gawer (ed). *Platforms, Markets and Innovation*, EE Publishing. Available at: https://doi.org/10.4337/9781849803311.00008

Bassan, S. (2025) 'Transparency ≠ Accountability? Rethinking Voluntary Vs. Mandatory Content Moderation Reports'. *SSRN Electronic Journal.* Available at: https://doi.org/10.2139/ssrn.5143075.

Bayer, J. (2024) 'The place of content ranking algorithms on the AI risk spectrum', *Telecommunications Policy,* 48(5), 102741. Available at: https://doi.org/10.1016/j.telpol.2024.102741.

Ben-David, A. (2020) 'Counter-archiving Facebook', *European Journal of Communication,* 35(3), 249-264. Available at: https://doi.org/10.1177/0267323120922069.

Bietti, E. (2020). 'From ethics washing to ethics bashing : A view on tech ethics from within moral philosophy', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 210-219. Available at: https://doi.org/10.1145/3351095.3372860.

Bouchaud, P., & Liénard, J. F. (2024) 'Beyond the guidelines : Assessing meta's political ad moderation in the EU, *Proceedings of the 2024 ACM on Internet Measurement Conference*, pp. 480-487. Available at: https://doi.org/10.1145/3646547.3689020

Bovens, M. (2007) 'Analysing and Assessing Accountability: A Conceptual Framework', *European Law Journal*, 13(4), pp. 447–468. Available at: https://doi.org/10.1111/j.1468-0386.2007.00378.x.

Carr, C. T. and Hayes, R. A. (2015) 'Social Media: Defining, Developing, and Divining', *Atlantic Journal of Communication*, 23(1), pp. 46–65. Available at: https://doi.org/10.1080/15456870.2015.972282.

Castaño-Pulgarín, S.A., Suárez-Betancur, N., Tilano Vega, L.M. and Herrera López, H.M. (2021) 'Internet, social media and online hate speech. Systematic review', *Aggression and Violent Behavior*, 58, p. 101608. Available at: https://doi.org/10.1016/j.avb.2021.101608.

Castets-Renard, C. (2020) 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement', *University of Illinois Journal of Law, Technology & Policy*, 2020(2), pp. 283–324.

Chander, A. and Krishnamurthy, V. (2018) ‘The Myth of Platform Neutrality’, *Georgetown law technology review,* 2(2), pp. 400-416.

Citron, D.K. (2018) ‘Extremist Speech, Compelled Conformity, and Censorship Creep’, *Notre Dame Law Review*, 93(3), p. 1035. Available at: https://scholarship.law.nd.edu/ndlr/vol93/iss3/3

Citron, D.K. and Penney, J. (2024) ‘Empowering Speech by Moderating It’, *Daedalus*, 153(3), pp. 31–44. Available at: https://doi.org/10.1162/daed_a_02087.

Coglianese, C. (2009) ‘The transparency president? The Obama administration and open government’, *Governance*, 22(4), 529‑544. Available at: https://doi.org/10.1111/j.1468-0491.2009.01451.x

Cook, C.L., Patel, A. and Wohn, D.Y. (2021) ‘Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms’, *Frontiers in Human Dynamics*, 3. Available at: https://doi.org/10.3389/fhumd.2021.626409.

Crogan, P. and Kinsley, S. (2012) ‘Paying Attention: Towards a Critique of the Attention Economy’, *Culture Machine*, 13. Available at: https://uwe-repository.worktribe.com/output/945724

Davis, M. and Xiao, J. (2021) ‘De-westernizing platform studies: history and logics of chinese and u. S. Platforms’, *International Journal of Communication*, 15(0), pp. 103-122. Available at: https://ijoc.org/index.php/ijoc/article/view/13961.

De Gregorio, G. (2020) ‘Democratising online content moderation: A constitutional framework’, *Computer Law & Security Review*, 36, p. 105374. Available at: https://doi.org/10.1016/j.clsr.2019.105374.

Douek, E. (2022) ‘Content Moderation as Systems Thinking’, *Harvard Law Review*, 136(2), pp. 526–607. Available at: https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/

Douglas-Scott, S. (1999) ‘The Hatefulness of Protected Speech: A Comparison of the American and European Approaches’, *William & Mary Bill of Rights Journal*, 7(2), pp. 305–346. Available at: https://scholarship.law.wm.edu/wmborj/vol7/iss2/2

Drolsbach, C. and Pröllochs, N. (2023) ‘Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database’ [PrePrint]. arXiv. Available at: https://doi.org/10.48550/arXiv.2312.04431.

Duivenvoorde, B. and Goanta, C. (2023) ‘The regulation of digital advertising under the DSA: A critical assessment’, *Computer Law & Security Review,* 51, p. 105870. Available at: https://doi.org/10.1016/j.clsr.2023.105870.

Duquenoy, P. (2005) ‘Ethics of Computing’, in J. Berleur and C. Avgerou (eds) *Perspectives and Policies on ICT in Society*. IFIP International Federation for Information Processing, 179, Springer, Boston, MA.
Available at: https://doi.org/10.1007/0-387-25588-5_12

Eder, N. (2024) ‘Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation’, *German Law Journal*, 25(7), pp. 1197–1218. Available at: https://doi.org/10.1017/glj.2024.24.

Farrand, B. (2023) ‘“Is This a Hate Speech?” The Difficulty in Combating Radicalisation in Coded Communications on Social media Platforms’, *European Journal on Criminal Policy and Research*, 29(3), pp. 477–493.
Available at: https://doi.org/10.1007/s10610-023-09543-z.

Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A. (2020) ‘Towards Transparency by Design for Artificial Intelligence’, *Science and Engineering Ethics*, 26(6), pp. 3333–3361. Available at: https://doi.org/10.1007/s11948-020-00276-4.

Flew, T. and Gillett, R. (2020) ‘Platform Policy - Evaluating Different Responses to the Challenges of Platform Power’, in *International Association for Media and Communication Research (IAMCR) annual conference*, Tampere, Finland.

Forssbæck, J. and Oxelheim, L. (2014) ‘The Multifaceted Concept of Transparency’, in J. Forssbæck and L. Oxelheim (eds) *The Oxford Handbook of Economic and Institutional Transparency*. Oxford University Press, pp. 3-30.
Available at: https://doi.org/10.1093/oxfordhb/9780199917693.013.0001.

Fox, J. (2007) ‘The Uncertain Relationship between Transparency and Accountability’, *Development in Practice*, 17(4/5), pp. 663–671.
Available at: https://doi.org/10.1080/09614520701469955

Frosio, G. (2023) ‘Platform responsibility in the digital services act: constitutionalising, regulating and governing private ordering’, in A. Savin and J. Trzakowski (eds.) *Research Handbook on EU Internet Law*. 2nd edition. EE Elgar.

Gillespie, T. (2010) ‘The politics of ‘platforms’’. *New Media & Society*, 12(3), 347-364.
Available at: https://doi.org/10.1177/1461444809342738

Gillespie, T. (2015). ‘Platforms Intervene’. *Social Media + Society*, 1(1).
Available at: https://doi.org/10.1177/2056305115580479

Gillespie, T. (2018) *Custodians of the Internet. Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press: New Haven & London.

Gorwa, R. and Ash, T.G. (2020) 'Democratic Transparency in the Platform Society', in N. Persily and J.A. Tucker (eds.) *Social Media and Democracy*. Cambridge: Cambridge University Press (SSRC Anxieties of Democracy), pp. 286–312. Available at: https://doi.org/10.1017/9781108890960

Gorwa, R., Binns, R. and Katzenbach, C. (2020) 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', *Big Data & Society*, 7(1), p. 2053951719897945. Available at: https://doi.org/10.1177/2053951719897945.

Grimmelmann, J. (2015) 'The Virtues of Moderation', *Yale Journal of Law & Technology*, 17, pp. 42-109. Available at: https://scholarship.law.cornell.edu/facpub/1486.

Guesmi, M., Chatti, M. A., Joarder, S., Ain, Q. U., Siepmann, C., Ghanbarzadeh, H., & Alatrash, R. (2023) 'Justification vs. Transparency: Why and How Visual Explanations in a Scientific Literature Recommender System', *Information*, 14(7), 401. Available at: https://doi.org/10.3390/info14070401

Hallinan, B., Scharlach, R. and Shifman, L. (2022) 'Beyond neutrality: conceptualizing platform values', *Communication Theory*, 32(2), pp. 201–222. Available at: https://doi.org/10.1093/ct/qtab008.

Haque, A. B., Islam, N., & Mikalef, P. (2024) 'To explain or not to explain : An empirical investigation of ai-based recommendations on social media platforms', *Electronic Markets*, 35(1), 2. Available at: https://doi.org/10.1007/s12525-024-00741-z.

Heald, D. (2006) 'Varieties of Transparency'. In C. Hood and D. Heald (eds). *Transparency: The Key to Better Governance?* Oxford University Press (Proceedings of the British Academy, 135), pp. 25–43.

Helberger, N., Pierson, J. and Poell, T. (2018) 'Governing online platforms: From contested to cooperative responsibility', *The Information Society*, 34(1), pp. 1–14. Available at: https://doi.org/10.1080/01972243.2017.1391913.

Husovec, M. (2023) 'Rising above liability: The Digital Services Act as a blueprint for the second generation of global internet rules', *Berkeley Technology Law Journal,* 38, pp. 101-137.

Husovec, M. (2024) 'The Digital Services Act's red line: what the Commission can and cannot do about disinformation', *Journal of Media Law*, 16(1), pp. 47–56. Available at: https://doi.org/10.1080/17577632.2024.2362483.

Imana, B., Korolova, A. and Heidemann, J. (2021) 'Auditing for Discrimination in Algorithms Delivering Job Ads', in *Proceedings of the Web Conference 2021.* New York, NY, USA: Association for Computing Machinery (WWW '21), pp. 3767–3778. Available at: https://doi.org/10.1145/3442381.3450077.

Izyumenko, E. et al. (2024) 'Online behavioural advertising, consumer empowerment and fair competition: Are the DSA transparency obligations the right answer?' *Rochester, NY: Social Science Research Network.* Available at: https://doi.org/10.2139/ssrn.4729118.

Jiang, J.A., Nie, P., Brubaker, J.R., and Fiesler, C. (2023) 'A Trade-off-centered Framework of Content Moderation', *ACM Transactions on Computer-Human Interaction*, 30(1), 3. Available at: https://doi.org/10.1145/3534929.

Jost, P., Kruschinski, S., Sülflow, M., Haßler, J. and Maurer, M. (2023) 'Invisible transparency: How different types of ad disclaimers on Facebook affect whether and how digital political advertising is perceived', *Policy & Internet*, 15(2), pp. 204–222. Available at: https://doi.org/10.1002/poi3.333.

Kabali, H.K., Irigoyen, M.M., Nunez-Davis, R., Budacki, J.G., Mohanty, S.H., Leister, K.P. and Bonner, R. (2015) 'Exposure and Use of Mobile Media Devices by Young Children', *Pediatrics*, 136(6), pp. 1044–1050. Available at: https://doi.org/10.1542/peds.2015-2151.

Karanicolas, M. (2021) 'A FOIA for Facebook: Meaningful Transparency for Online Platforms', *Saint Louis University Law Journal*, 66(1), pp. 49–78. Available at: https://scholarship.law.slu.edu/lj/vol66/iss1/4/

Kaushal, R., Van De Kerkhof, J., Goanta, C., Spanakis, G., and Iamnitchi, A. (2024) 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *Proceedings of the FAccT '24: 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro Brazil: ACM, pp. 1121–1132. Available at: https://doi.org/10.1145/3630106.3658960.

Kim, H. and Lee, T. H. (2018) 'Strategic CSR Communication: A Moderating Role of Transparency in Trust Building', *International Journal of Strategic Communication*, 12(2), pp. 107–124. Available at: https://doi.org/10.1080/1553118X.2018.1425692.

Koivisto, I. (2019) 'Towards Critical Transparency Studies. Emmanuel Alloa and Dieter Thomä (eds): Transparency, Society and Subjectivity: Critical Perspectives. Palgrave Macmillan 2018, 408 pp', *Res Publica,* 25, pp. 439-443. Available at: https://doi.org/10.1007/s11158-019-09425-4

Kosters, L. and Gstrein, O.J. (2024) ‘TikTok and Transparency Obligations in the EU Digital Services Act (DSA) – A Scoping Review’, *Zeitschrift für europarechtliche Studien*, 27(1), pp. 110–145. Available at: https://doi.org/10.5771/1435-439X-2024-1-110.

Langley, P. and Leyshon, A. (2017) ‘Platform capitalism: The intermediation and capitalisation of digital economic circulation’, *Finance and Society*, 3(1), pp. 11-31. Available at: https://doi.org/10.2218/finsoc.v3i1.1936

Langvardt, K. (2018) ‘Regulating Online Content Moderation’, *Georgetown Law Journal*, 106(5), pp. 1353–1388. Available at: https://ssrn.com/abstract=3024739

Lee, A. and Chung, T.-L.D. (2023) ‘Transparency in corporate social responsibility communication on social media’, *International Journal of Retail & Distribution Management*, 51(5), pp. 590–610. Available at: https://doi.org/10.1108/IJRDM-01-2022-0038.

Leerssen, P. (2020) ‘The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems’, *European Journal of Law and Technology*, 11(2). Available at: https://ejlt.org/index.php/ejlt/article/view/786.

Leerssen, P. (2023) ‘An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation’, *Computer Law & Security Review*, 48, p. 105790. Available at: https://doi.org/10.1016/j.clsr.2023.105790.

Le Merrer, E., Morgan, B., & Trédan, G. (2021) ‘Setting the record straighter on shadow banning’, *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications,* pp. 1-10. Available at: https://doi.org/10.1109/INFOCOM42981.2021.9488792.

Leone de Castris, A. (2024) ‘Types of Platform Transparency: An Analysis of Discourse Around Transparency and Global Digital Platforms’, *Public Integrity,* pp. 1–15. Available at: https://doi.org/10.1080/10999922.2024.2304741.

Lessig, L. (1999) *Code: And Other Laws Of Cyberspace*. New York: Basic Books.

Lewandowsky, S., Robertson, R.E. and DiResta, R. (2024) ‘Challenges in Understanding Human-Algorithm Entanglement During Online Information Consumption’, *Perspectives on Psychological Science*, 19(5), pp. 758–766. Available at: https://doi.org/10.1177/17456916231180809.

Lodder, A.R. (2017) ‘European Union E-Commerce Directive - Article by Article Comments’. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=1009945

Maroni, M. (2023) '"Mediated transparency": The Digital Services Act and the legitimisation of platform power', in M, Hillebrandt, P. Leino-Sandberg, and I. Koivisto (eds.) (In)visible European Government. Routledge, pp. 305-327. Available at: https://doi.org/10.4324/9781003257936

Matheus, R. and Janssen, M. (2019) 'A Systematic Literature Study to Unravel Transparency Enabled by Open Government Data: The Window Theory', *Public Performance & Management Review*, 43(3), pp. 503–534. Available at: https://doi.org/10.1080/15309576.2019.1691025.

McKee, D. (2017) 'The platform economy: natural, neutral, consensual and efficient?', *Transnational Legal Theory*, 8(4), pp. 455–495. Available at: https://doi.org/10.1080/20414005.2017.1416516.

McMillan Cottom, T. (2020). 'Where Platform Capitalism and Racial Capitalism Meet: The Sociology of Race and Racism in the Digital Society'. *Sociology of Race and Ethnicity*, 6(4), 441-449. https://doi.org/10.1177/2332649220949473

Meijer, A. (2014) 'Transparency', in M. Bovens, R. Goodin, and T. Schillemans (eds.) *The Oxford Handbook of Public Accountability*. Oxford University Press, pp. 508-524. Available at: https://doi.org/10.1093/oxfordhb/9780199641253.013.0043.

Mejias, U.A. and Couldry, N. (2019) 'Datafication', *Internet Policy Review*, 8(4). Available at: https://doi.org/10.14763/2019.4.1428

Minihold, S., & Votta, F. (2024) 'Accepting exclusion : Examining the (Un)intended consequences of data-driven campaigns', *Media and Communication*, 12(0). Available at: https://doi.org/10.17645/mac.8685.

Mitova, E., Blassnig, S., Strikovic, E., Urman, A., de Vreese, C., & Esser, F. (2023) 'Exploring users' desire for transparency and control in news recommender systems: A five-nation study', *Journalism*, 25(10), pp. 2001-2021. Available at: https://doi.org/10.1177/14648849231222099.

Nannini, L. et al. (2024) 'Beyond phase-in: assessing impacts on disinformation of the EU Digital Services Act', *AI and Ethics* [Preprint]. Available at: https://doi.org/10.1007/s43681-024-00467-w.

Naurin, D. (2007) 'Transparency, Publicity, Accountability – The missing links'. *CONNEX-RG 2 workshop on "Delegation and Mechanisms of Accountability in the EU"*, Uppsala.

Newman, J.M. (2019) 'Antitrust in Digital Markets', *Vanderbilt Law Review*, 72(5), pp. 1497-1561. Available at: https://scholarship.law.vanderbilt.edu/vlr/vol72/iss5/2/

Njie, B. and Asimiran, S. (2014) ‘Case Study as a Choice in Qualitative Methodology’, *IOSR Journal of Research & Method in Education*, 4(3), pp. 35-40.

Nourooz-Pour, H. (2024) ‘Voices and values: the challenging odyssey of Meta to harmonize human rights with content moderation’, *International Journal of Law and Information Technology*, 32(1), p. eaae009. Available at: https://doi.org/10.1093/ijlit/eaae009.

Nunziato, D. (2023). The digital services act and the brussels effect on platform content moderation. *Chicago Journal of International Law*, 24(1). Available at: https://chicagounbound.uchicago.edu/cjil/vol24/iss1/6

Olesen, T. (2025) ‘Big Tech whistleblowing: Frances Haugen and the Facebook Files’, *Organization*, 0(0), pp. 1-20. Available at: https://doi.org/10.1177/13505084251321785.

Padfield, R., Matoh, S., Tyson, A., Wong, C., Bridge, G., & Dales, A. (2025). Transparency or map‐washing? Digital geospatial visualisation tools in the palm oil industry. *Business Strategy and the Environment*, bse.4280. Available at: https://doi.org/10.1002/bse.4280

Palmeira Ferraz, T.P., Dias Duarte, C.H., Ribeiro, M.F., Braga Takayanagi, G.G., Alcoforado, A., de Deus Lopes, R. and Susi, M. (2024) ‘Explainable AI to Mitigate the Lack of Transparency and Legitimacy in Internet Moderation’, *Estudos Avançados*, 38, pp. 381–405. Available at: https://doi.org/10.1590/s0103-4014.202438111.020.

Palumbo, A. (2024) ‘A Medley of Public and Private Power in DSA Content Moderation for Harmful but Legal Content: An Account of Transparency, Accountability and Redress Challenges’, *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 15(3), pp. 246–268.

Pappas, C., Argyraki, K., Bechtold, S. and Perrig, A. (2015) ‘Transparency instead of neutrality’, *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*, Philadelphia PA USA: ACM, pp. 1–7. Available at: https://doi.org/10.1145/2834050.2834082.

Poell, T., Nieborg, D. and van Dijck, J. (2019) ‘Platformisation’, *Internet Policy Review*, 8(4). Available at: https://doi.org/10.14763/2019.4.1425.

Portaru, A. (2017) ‘Freedom of Expression Online: The Code of Conduct on Countering Illegal Hate Speech Online Doctrine’, *Revista Romana de Drept European*, 2017(4), pp. 77–91.

Price, M.E. and Price, J.M. (2023) ‘Building Legitimacy in the Absence of the State: Reflections on the Facebook Oversight Board’, *International Journal of Communication*, 17(2023), pp. 3315–3325. Available at: https://ijoc.org/index.php/ijoc/article/view/19915

Relia, K., Li, Z., Cook, S.H. and Chunara, R. (2019) ‘Race, Ethnicity and National Origin-Based Discrimination in Social Media and Hate Crimes across 100 U.S. Cities’, *Proceedings of the International AAAI Conference on Web and Social Media*, 13, pp. 417–427. Available at: https://doi.org/10.1609/icwsm.v13i01.3354.

Reid, A., Pendleton, S.M. and Czabovsky, L.E.H.J. (2024a) ‘Social Media Transparency Reports: Longitudinal Content Analysis of News Coverage’, *The Journal of Social Media in Society*, 13(1), pp. 122-154. Available at: https://doi.org/10.2139/ssrn.4891917.

Reid, A., Ringel, E. and Pendleton, S.M. (2024b) ‘Transparency reports as CSR reports: motives, stakeholders, and strategies’, *Social Responsibility Journal*, 20(1), pp. 81–107. Available at: https://doi.org/10.1108/SRJ-03-2023-0134.

Reid, A. and Ringel, E. (2025) ‘Digital intermediaries and transparency reports as strategic communications’, *The Information Society*, 41(2), pp. 91–109. Available at: https://doi.org/10.1080/01972243.2025.2453529.

Rieder, B. and Hofmann, J. (2020) ‘Towards platform observability’, *Internet Policy Review*, 9(4). Available at: https://doi.org/10.14763/2020.4.1535.

Roberts, S.T. (2018) ‘Digital detritus: “Error” and the logic of opacity in social media content moderation’, *First Monday*, 23(3). Available at: https://doi.org/10.5210/fm.v23i3.8283.

Sander, B. (2020) ‘Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation’, *Fordham International Law Journal*, 43(4), p. 939. Available at: https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3

Savin, A. and Trzakowski, J. (2023) *Research handbook on EU Internet Law*. 2nd edition. EE Elgar.

Schlag, G. (2023) ‘European Union’s Regulating of Social Media: A Discourse Analysis of the Digital Services Act’, *Politics and Governance*, 11(3). Available at: https://doi.org/10.17645/pag.v11i3.6735.

Schauer, F. (2011) ‘Transparency in three dimensions’, *University of Illinois Law Review*, 2011(4), pp. 1339-1357.

Schwemer, S.F. (2023) ‘Digital Services Act: a reform of the e-Commerce Directive and much more’, in A. Savin and J. Trzakowski (eds.) *Research Handbook on EU Internet Law*. 2nd edition. EE Elgar.

Siegel, A.A. (2020) ‘Online Hate Speech’, in J.A. Tucker and N. Persily (eds) *Social Media and Democracy*. Cambridge: Cambridge University Press (SSRC Anxieties of Democracy), pp. 56–88. Available at: https://doi.org/10.1017/9781108890960

Söderlund, K., Engström, E., Haresamudram, K., Larsson, S. and Strimling, P. (2024) ‘Regulating high-reach AI: On transparency directions in the Digital Services Act’, *Internet Policy Review*, 13(1). Available at: https://doi.org/10.14763/2024.1.1746

Spindler, G. (2023) ‘EU Internet policy in the 2020s’, in A. Savin and J. Trzakowski (eds.) *Research Handbook on EU Internet Law*. 2nd edition. EE Elgar, pp. 2–45.

Srnicek, N. (2016) *Platform Capitalism*. Newark, United Kingdom: Polity Press.

Ștefăniță, O. and Buf, D.-M. (2021) ‘Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research’, *Romanian Journal of Communication and Public Relations*, 23(1), pp. 47–55.
Available at: https://doi.org/10.21018/rjcpr.2021.1.322.

Storms, E., Alvarado, O., & Monteiro-Krebs, L. (2022) ‘« Transparency is meant for control » and vice versa : Learning from co-designing and evaluating algorithmic news recommenders’, *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), pp. 1-24. Available at: https://doi.org/10.1145/3555130.

Suzor, N. (2018) ‘Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms’, *Social Media + Society*, 4(3), p. 2056305118787812. Available at: https://doi.org/10.1177/2056305118787812.

Suzor, N.P., Myers West, S., Quodling, A. and York, J. (2019) ‘What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation’, *International Journal of Communication*, 13(0), p. 18.

Trujillo, A., Fagni, T. and Cresci, S. (2024) ‘The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media’ [PrePrint]. arXiv. Available at: https://doi.org/10.48550/ARXIV.2312.10269.

Turillazzi, A. et al. (2023) ‘The digital services act: an analysis of its ethical, legal, and social implications’, *Law, Innovation and Technology*, 15(1), pp. 83–106.
Available at: https://doi.org/10.1080/17579961.2023.2184136.

Tyler, T.R., Meares, T.L. and Katsaros, M. (2025) 'New Worlds Arise: Online Trust and Safety', *Annual Review of Criminology*, 8, pp. 171–192. Available at: https://doi.org/10.1146/annurev-criminol-111523-122337.

Uras, B.R. (2020) 'Finance and development: Rethinking the role of financial transparency', *Journal of Banking & Finance*, 111, p. 105721. Available at: https://doi.org/10.1016/j.jbankfin.2019.105721.

Urman, A. and Makhortykh, M. (2023) 'How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms', *Telecommunications Policy*, 47(3), p. 102477. Available at: https://doi.org/10.1016/j.telpol.2022.102477.

Vaccaro, K., Sandvig, C. and Karahalios, K. (2020) '"At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation', *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 167. Available at: https://doi.org/10.1145/3415238.

Vaccaro, K., Xiao, Z., Hamilton, K. and Karahalios, K. (2021) 'Contestability For Content Moderation', *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 318. Available at: https://doi.org/10.1145/3476059.

Van De Kerkhof, J. and Goanta, C. (2024) 'Shadowbanned on X: The DSA in Action', *European Journal of Risk Regulation*, pp. 1–8. Available at: https://doi.org/10.1017/err.2024.81.

Veltri, G.A., Lupiáñez-Villanueva, F., Folkvord, F., Theben, A. and Gaskell, G. (2023) 'The impact of online platform transparency of information on consumers' choices', *Behavioural Public Policy,* 7(1), pp. 55–82. Available at: https://doi.org/10.1017/bpp.2020.11.

Votta, F., Dobber, T., Guinaudeau, B., Helberger, N., & de Vreese, C. (2024) 'The Cost of Reach: Testing the Role of Ad Delivery Algorithms in Online Political Campaigns', *Political Communication*, pp. 1–33. Available at: https://doi.org/10.1080/10584609.2024.2439317

Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J. and Singh, J. (2020) 'Regulating transparency? Facebook, Twitter and the German Network Enforcement Act', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain, pp. 261–271. Available at: https://doi.org/10.1145/3351095.3372856.

Walters, M.A., Paterson, J., Brown, R. and McDonnell, L. (2020) ‘Hate Crimes Against Trans People: Assessing Emotions, Behaviors, and Attitudes Toward Criminal Justice Agencies’, *Journal of Interpersonal Violence*, 35(21–22), pp. 4583–4613. Available at: https://doi.org/10.1177/0886260517715026.

Wang, S., Zhang, X., Wang, Y., & Ricci, F. (2024) ‘Trustworthy recommender systems’, *ACM Transactions on Intelligent Systems and Technology*, 15(4), pp. 1-20. Available at: https://doi.org/10.1145/3627826.

Ward, S. (2017) ‘From fontainebleau to facebook: the early modern discourse of personal sincerity and its echoes in the contemporary discourse of organisational transparency’, *Systems Research and Behavioral Science*, 34(2), pp. 139–147. Available at: https://doi.org/10.1002/sres.2448.

Williams, J. (2023) ‘Deplatforming sex education on Meta: sex, power, and content moderation’, *Media International Australia*, 194(1), 36-52. Available at: https://doi.org/10.1177/1329878X231210612.

Wilson, R.A. and Land, M.K. (2021) ‘Hate Speech on Social Media: Content Moderation in Context’, *Connecticut Law Review*, 52(3), pp. 1029–1076. Available at: https://digitalcommons.lib.uconn.edu/law_review/449.

Wood, A. (2021) ‘Learning from campaign finance information’, *Emory Law Journal*, 70(5), p. 1091-1142. Available at: https://scholarlycommons.law.emory.edu/elj/vol70/iss5/2.

Wu, T. (2019) ‘Blind Spot: The Attention Economy and the Law’, *Antitrust Law Journal*, 82, pp.771-806.

Xiao, L. Y. (2025) ‘Illegal loot box advertising on social media? An empirical study using the Meta and TikTok ad transparency repositories’, *Computer Law & Security Review*, 56, 106069. Available at: https://doi.org/10.1016/j.clsr.2024.106069

Xue, C., Tian, W. and Zhao, X. (2020) ‘The Literature Review of Platform Economy’, *Scientific Programming,* 2020(1), p. 8877128. Available at: https://doi.org/10.1155/2020/8877128.

Yu, P.K. (2021) ‘Beyond Transparency and Accountability: Three Additional Features Algorithm Designers Should Build into Intelligent Platforms’, *Northeastern University Law Review*, 13(1), pp. 263–296.

Zalnieriute, M. (2021) ‘“Transparency Washing” in the Digital Age: A Corporate Agenda of Procedural Fetishism Special Issue: Transparency in the Digital Environment’, *Critical Analysis of Law: An International & Interdisciplinary Law Review*, 8(1), pp. 139–153.

Zeng, J. and Kaye, D.B.V. (2022) 'From content moderation to visibility moderation: A case study of platform governance on TikTok', *Policy & Internet*, 14(1), pp. 79–95. Available at: https://doi.org/10.1002/poi3.287.

Zuboff, S. (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power,* New York: PublicAffairs.

Zuckerman, E. and Rajendra-Nicolucci, C. (2023) 'From Community Governance to Customer Service and Back Again: Re-Examining Pre-Web Models of Online Governance to Address Platforms' Crisis of Legitimacy', *Social Media + Society*, 9(3), p. 20563051231196864.
Available at: https://doi.org/10.1177/20563051231196864.

Websites, newspaper articles and reports

Algorithm Watch (2024, November 29) '*DSA: Erste Risikobewertungsberichte über systemische Risiken von großen Online-Plattformen lassen viele Fragen offen*'. Available at: https://algorithmwatch.org/de/dsa-risikobewertungsberichte/ (Accessed April 15, 2025)

Bell, K. (2018, April 24) '*Facebook just made a major change to how it polices content*', Mashable. Available at: https://mashable.com/article/facebook-new-community-standards-appeals (Accessed April 15, 2025)

Bernard, T. (2024, December 20) '*Reading the Systemic Risk Assessments for Major Speech Platforms: Notes and Observations*', Tech Policy Press. Available at: https://www.techpolicy.press/reading-the-systemic-risk-assessments-for-major-speech-platforms-notes-and-observations/ (Accessed April 15, 2025)

Bickert, M. (2018, April 24) *'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process'*, Meta Newsroom. Available at: https://about.fb.com/news/2018/04/comprehensive-community-standards/ (Accessed April 15, 2025)

Bradford, B., Grisel, F., Meares, T.L., Owens, E., Pineda, B.L., Shapiro, J.N., Tyler, T.R. and Peterman, D.E. (2019) *Report Of The Facebook Data Transparency Advisory Group*. The Justice Collaboratory, Yale Law School.

Breton, T. (2022, January 19) '*Speech by Commissioner Breton on the Digital Services Act'*, European Commission. Available at: https://ec.europa.eu/commission/presscorner/detail/en/speech_22_431 (Accessed April 15, 2025)

Broxmeyer, J. (2021, July 15) '*Facebook's First Quarterly Update on the Oversight Board*', Meta Newsroom. Available at: https://about.fb.com/news/2021/07/facebooks-first-quarterly-update-on-the-oversight-board/ (Accessed April 15, 2025)

Bouchaud, P. (2025) '*Pay-to-Play: Meta's Community (double) Standards on Pornographic Ads*', AI Forensics. Available at: https://aiforensics.org/work/meta-porn-ads

Buri, I. and Van Hoboken, J. (2021) *'The Digital Services Act (DSA) proposal: a critical overview'*, Discussion paper, Digital Services Act (DSA) Observatory, Institute for Information Law (IViR), University of Amsterdam.

Campetti Amaral, R., Trigo, M., Rebello Pereira, F., and Salomão Jabra, A. (2024, September 04) '*Brazil: National Consumer Secretariat establishes transparency and data quality criteria for digital platforms*', Global Compliance News. Available at: https://www.globalcompliancenews.com/2024/09/04/brazil-national-consumer-secretariat-establishes-transparency-and-data-quality-criteria-for-digital-platforms/ (Accessed April 15, 2025)

Caplan, R. (2018) *'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches'*, Data & Society. Available at: https://datasociety.net/library/content-or-context-moderation/

Chakrabarti, S. (2018, January 22) '*Hard Questions: What Effect Does Social Media Have on Democracy?'*, Meta Newsroom. Available at: https://about.fb.com/news/2018/01/effect-social-media-democracy/ (Accessed April 15, 2025)

Clegg, N. (2023, August 22) '*New Features and Additional Transparency Measures as the Digital Services Act Comes Into Effect'*, Meta Newsroom. Available at: https://about.fb.com/news/2023/08/new-features-and-additional-transparency-measures-as-the-digital-services-act-comes-into-effect/ (Accessed April 15, 2025)

European Commission (n.d.a) '*Overview Documentation*', DSA Transparency Database. Available at: https://transparency.dsa.ec.europa.eu/page/documentation (Accessed April 15, 2025)

European Commission (n.d.b) '*Welcome to the DSA Transparency Database!*', DSA Transparency Database. Available at: https://transparency.dsa.ec.europa.eu/ (Accessed April 15, 2025)

European Commission (2019, September 27) *'Assessment of the Code of Conduct on Hate Speech on line- State of Play'*, Information Note 12522/19. Available at: https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (Accessed April 15, 2025)

European Commission (2024, October 28) *'Draft Regulation - COMMISSION DELEGATED REGULATION (EU) supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by laying down the technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data pursuant to Article 40 of Regulation (EU) 2022/2065'*, Draft Ares(2024)7652659. Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act_en

European Commission (2025, February 06) '*Supervision of the designated very large online platforms and search engines under DSA'*. Available at: https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses (Accessed April 15, 2025)

European Commission - Directorate-General for Communications Networks, Content and Technology (DG CONNECT) (2023) 'Digital Services Act Transparency Database'. Available at: https://doi.org/10.2906/134353607485211

EY (2024) '*Independent Audit on Facebook For the Period of 29 August 2023 to 30 June 2024 With an Assurance Report of Independent Accountants regarding Regulation (EU) 2022/2065, the Digital Services Act (DSA)*'.

Facebook (2011) '*Facebook Community Standards*' in WebArchive, September 29, 2011. Available at: https://web.archive.org/web/20110929191437/https://www.facebook.com/communitystandards (Accessed April 15, 2025)

Facebook (2014) '*Facebook Community Standards'* in WebArchive, July 01, 2014. Available at: https://web.archive.org/web/20140701143118/https://www.facebook.com/communitystandards (Accessed April 15, 2025)

Facebook (2020) '*14. Adult Nudity and Sexual Activity'* in WebArchive, December 21, 2020. Available at: https://web.archive.org/web/20201221155333/https://www.facebook.com/communitystandards/adult_nudity_sexual_activity (Accessed April 15, 2025)

Frenkel, S. and Benner, K. (2018, February 17) '*To Stir Discord in 2016, Russians Turned Most Often to Facebook'*, The New York Times. Available at: https://www.nytimes.com/2018/02/17/technology/indictment-russian-tech-facebook.html (Accessed April 15, 2025)

Gebhart, G. (2018, May 31) '*Who Has Your Back? Censorship Edition 2018'*, Electronic Frontier Foundation. Available at: https://www.eff.org/who-has-your-back-2018 (Accessed April 15, 2025)

Gebhart, G. (2019, June 12) '*Who Has Your Back? Censorship Edition 2019*, Electronic Frontier Foundation. Available at: https://www.eff.org/wp/who-has-your-back-2019 (Accessed April 15, 2025)

Gerken, T. (2018, March 21) '*WhatsApp co-founder says it is time to delete Faceboo*k', BBC. Available at: https://www.bbc.com/news/blogs-trending-43470837 (Accessed April 15, 2025)

Han, K. (2016, July 07) 'In which I'm blocked from Facebook for… what?' Medium. Available at: https://medium.com/@kixes/in-which-im-blocked-from-facebook-for-what-37e6608c38c6 (Accessed April 15, 2025)

Hern, A. (2021, February 11) '*Facebook moderators 'told not to discuss working conditions'*, The Guardian. Available at: https://www.theguardian.com/technology/2021/feb/11/facebook-moderators-say-they-were-told-not-to-discuss-covid-working-conditions (Accessed April 15, 2025)

Hutchinson, A. (2019, February 12) '*Instagram Adds New Warnings for Accounts Which Are Close to Being Banned*', Social Media Today. Available at: https://www.socialmediatoday.com/news/instagram-adds-new-warnings-for-accounts-which-are-close-to-being-banned/559079/ (Accessed April 15, 2025)

Hutchinson, A. (2020, February 12) '*Instagram Launches New Appeals Process for Disabled Accounts, Adds Report Tracking In-App*', Social Media Today. Available at: https://www.socialmediatoday.com/news/instagram-launches-new-appeals-process-for-disabled-accounts-adds-report-t/572122/ (Accessed April 15, 2025)

Hutchinson, A. (2023, August 2022) '*EU Users Can Soon Opt Out of Algorithmic Sorting on Facebook and Instagram*', Social Media Today. Available at: https://www.socialmediatoday.com/news/algorithmic-sorting-facebook-instagram-opt-out-EU/691567/ (Accessed April 15, 2024)

Hutchinson, A. (2025, January 29) *'Threads Reaches 320M Monthly Active Users'*, Social Media Today. Available at: https://www.socialmediatoday.com/news/threads-rises-to-320-million-active-users/738722 (Accessed April 15, 2025)

Levin, S. (2017, December 15) '*Facebook admits it poses mental health risk – but says using site more can help*', The Guardian. Available at: https://www.theguardian.com/technology/2017/dec/15/facebook-mental-health-psychology-social-media (Accessed April 15, 2025)

Luria, M. (2022) '*"This is Transparency to me" User Insights into Recommendation Algorithms Reporting'*, Center for Democracy and Technology. Available at: https://doi.org/10.31219/osf.io/qfcpx.

MacCarthy, M. (2022, November 1) *'Transparency is essential for effective social media regulation'*, Brookings. Available at: https://www.brookings.edu/articles/transparency-is-essential-for-effective-social-media-regulation/ (Accessed April 15, 2025)

'Machine Readable' (n.d.) *Open Data Handbook.* Available at: https://opendatahandbook.org/glossary/en/terms/machine-readable/ (Accessed April 15, 2025)

Madiega, T. (2022) *Digital Services Act.* Briefing PE 689.357. European Parliamentary Research Service.

Mansell, R., Durach, F., Kettemann, M., Lenoir, T. Procter, R., Tripathi, G., and Tucker, E. (2025) *Information ecosystems and troubled democracy. A Global Synthesis of the State of Knowledge on News Media, AI and Data Governance.* Forum on Information and Democracy. Available at: https://observatory.informationdemocracy.org/wp-content/uploads/2024/12/rapport_forum_information_democracy_2025.pdf.

Meta (n.d.a) *'Working at Meta means making every connection matter'*. Available at: https://www.metacareers.com/culture/ (Accessed April 15, 2025)

Meta (n.d.b) '*Exigences relatives au bénéficiaire et au payeur pour les publicités ciblant l'Union européenne*', Facebook. Available at: https://www.facebook.com/business/help/605021638170961 (Accessed April 15, 2025)

Meta (2018a, November 15) '*Product Policy Forum Minutes*', Meta Newsroom. Available at: https://about.fb.com/news/2018/11/content-standards-forum-minutes/ (Accessed April 15, 2025)

Meta (2018b, December 28) '*Facts About Content Review on Facebook*', Meta Newsroom. Available at: https://about.fb.com/news/2018/12/content-review-facts/ (Accessed April 15, 2025)

Meta (2019a, March 31) '*More Clarity, More Control*', Meta Newsroom. Available at: https://about.fb.com/news/2019/03/inside-feed-why-am-i-seeing-this-post/ (Accessed April 15, 2025)

Meta (2019b, March 31) '*Why Am I Seeing This? We Have an Answer for You*', Meta Newsroom. Available at: https://about.fb.com/news/2019/03/why-am-i-seeing-this/ (Accessed April 15, 2025)

Meta (2021) *'Widely Viewed Content Report: What People See on Facebook - Q2 2021'*. Available at: https://transparency.meta.com/data/widely-viewed-content-report/ (Accessed April 15, 2025)

Meta (2023a) '*Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook*'. Transparency Center. Available at: https://transparency.meta.com/sr/dsa-transparency-report-oct2023-facebook/

Meta (2023b, June 29) '*Introducing 22 system cards that explain how AI powers experiences on Facebook and Instagram*', Meta AI. Available at: https://ai.meta.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/ (Accessed April 15, 2025)

Meta (2023c, November 07) '*Content actioned*', Transparency Center. Available at: https://transparency.meta.com/policies/improving/content-actioned-metric/ (Accessed April 15, 2025)

Meta (2024a) *'Regulation (EU) 2022/2065 Digital Services Act (DSA) Systemic Risk Assessment and Mitigation Report for Facebook'*. Transparency Center. Available at: https://transparency.meta.com/reports/regulatory-transparency-reports/

Meta (2024b) *'Regulation (EU) 2022/2065 Digital Services Act (DSA) Systemic Risk Assessment and Mitigation Report for Instagram*. Transparency Center. Available at: https://transparency.meta.com/reports/regulatory-transparency-reports/

Meta (2024c) '*Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook*'. Transparency Center. Available at: https://transparency.meta.com/sr/dsa-transparency-report-apr2024-facebook

Meta (2024d) '*Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook*'. Transparency Center. Available at: https://transparency.meta.com/sr/dsa-transparency-report-sep2024-facebook

Meta (2024e) *'Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram'*. Transparency Center. Available at: https://transparency.meta.com/reports/regulatory-transparency-reports/

Meta (2024f) '*Widely Viewed Content Report: What People See on Facebook - Q4 2024'*, Transparency Center.
Available at: https://transparency.meta.com/data/widely-viewed-content-report/ (Accessed April 15, 2025)

Meta (2024g, November 12) *'How enforcement technology works'*, Transparency Center. Available at: https://transparency.meta.com/enforcement/detecting-violations/how-enforcement-technology-works/ (Accessed April 15, 2025)

Meta (2024h, November 12) *'Policy Forum Minutes'*, Transparency Center. Available at: https://transparency.meta.com/policies/improving/policy-forum-minutes/ (Accessed April 15, 2025)

Meta (2024i, November 12) '*Restricting accounts'*, Transparency Center. Available at: https://transparency.meta.com/enforcement/taking-action/restricting-accounts/ (Accessed April 15, 2025)

Meta (2025a, January 29) *'Meta Reports Fourth Quarter and Full Year 2024 Results'*. Available at: https://investor.atmeta.com/investor-news/press-release-details/2025/Meta-Reports-Fourth-Quarter-and-Full-Year-2024-Results/default.aspx (Accessed April 15, 2025)

Meta (2025b, February 19) '*How we apply our content policies*', Transparency Center. Available at: https://transparency.meta.com/enforcement/taking-action/applying-content-policies/ (Accessed April 15, 2025)

Meta (2025c, March 07) '*Instagram Explore AI system*', Transparency Center. Available at: https://transparency.meta.com/features/explaining-ranking/ig-explore/?referrer=1 (Accessed April 15, 2025)

Meta (2025d, March 24) '*Facebook Notifications AI system*', Transparency Center. Available at: https://transparency.meta.com/features/explaining-ranking/fb-notifications/?referrer=1 (Accessed April 15, 2025)

Neate, R. (2018, July 26) *'Over $119bn wiped off Facebook's market cap after growth shock'*, The Guardian. Available at: https://www.theguardian.com/technology/2018/jul/26/facebook-market-cap-falls-109bn-dollars-after-growth-shock (Accessed April 15, 2025)

Newton, C. (2017, December 15) *'Facebook says 'passively consuming' the News Feed will make you feel worse about yourself'*, The Verge. Available at: https://www.theverge.com/2017/12/15/16781448/facebook-makes-you-feel-bad-study-research (Accessed April 15, 2025)

Ouangari, L. (2021, November 10) '*« Facebook Files ». Frances Haugen interrogée par les députés : ce qu'a dit la lanceuse d'alerte'*, Ouest France. Available at: https://www.ouest-france.fr/high-tech/facebook/facebook-files-transparence-desinformation-frances-haugen-repond-aux-deputes-francais-b31250b4-4200-11ec-bf34-dccbc8c1efb2 (Accessed April 15, 2025)

Oversight Board (n.d.) *'How we do our work'*. Available at: https://www.oversightboard.com/our-work/ (Accessed April 15, 2025)

Oversight Board (2021, October 21) *'Oversight Board demands more transparency from Facebook'*. Available at: https://www.oversightboard.com/news/215139350722703-oversight-board-demands-more-transparency-from-facebook/ (Accessed April 15, 2025)

Pavón, P. (2023, February 14) '*Increasing Our Ads Transparency'*, Meta Newsroom. Available at: https://about.fb.com/news/2023/02/increasing-our-ads-transparency/ (Accessed April 15, 2025)

Perrigo, B. (2022, June 16) *Frances Haugen Calls for 'Solidarity' With Facebook Content Moderators in Conversation with Whistleblower Daniel Motaung*, Time. Available at: https://time.com/6188272/frances-haugen-daniel-motaung-facebook-whistleblowers/ (Accessed April 15, 2025)

Rankin, J. (2025, April 11) *'EU will not rip up tech rules for trade deal with Trump, senior official says'*, The Guardian. Available at: https://www.theguardian.com/world/2025/apr/11/eu-will-not-rip-up-tech-rules-for-trade-deal-with-trump-senior-official-says (Accessed April 15, 2025)

Reiff, N. (2024, July 30) *'Top Facebook (Meta) Shareholders'*, Investopedia. Available at: https://www.investopedia.com/articles/insights/082216/top-9-shareholders-facebook-fb.asp (Accessed April 15, 2025)

'Santa Clara Principles on Transparency and Accountability in Content Moderation' (2018). Available at: https://santaclaraprinciples.org/

Silver, E. (July 26, 2018) '*Hard Questions: Who Reviews Objectionable Content on Facebook — And Is the Company Doing Enough to Support Them?*', Meta Newsroom. Available at: https://about.fb.com/news/2018/07/hard-questions-content-reviewers/ (Accessed April 15, 2025)

Statista (2025, February 12) *'Classement des réseaux sociaux les plus populaires dans le monde en janvier 2025, selon le nombre d'utilisateurs actifs'*. Available at: https://fr.statista.com/statistiques/570930/reseaux-sociaux-mondiaux-classes-par-nombre-d-utilisateurs/ (Accessed April 15, 2025)

Sunstein, C.R. (2018, January 22) '*Cass R. Sunstein: Is Social Media Good or Bad for Democracy?*', Meta Newsroom. Available at: https://about.fb.com/news/2018/01/sunstein-democracy/ (Accessed April 15, 2025)

'Transparency' (n.d.) *Cambridge Dictionary*. Available at: https://dictionary.cambridge.org/dictionary/english/transparency (Accessed March 14, 2025).

Tsukayama, H. (2014, June 12) *'Facebook makes some big changes to its advertisements'*, The Washington Post. Available at: https://www.washingtonpost.com/news/the-switch/wp/2014/06/12/facebook-makes-some-big-changes-to-its-advertisements/ (Accessed April 15, 2025)

Urwin, R. (2025, March 12) '*Metamorphosis: One woman's journey through the Facebook looking glass*', The Press. Available at: https://www.thepress.co.nz/world-news/360617619/metamorphosis-one-womans-journey-through-facebook-looking-glass (Accessed April 15, 2025)

Van Zuylen-Wood, S. (2019, February 26) *'"Men Are Scum": Inside Facebook's War on Hate Speech'*, Vanity Fair. Available at: https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech (Accessed April 15, 2025)

WhatsApp (n.d.) '*About how WhatsApp recommends channels'*, Help Center. Available at: https://faq.whatsapp.com/962978635456336/?helpref=search&cms_id=962978635456336&search_session_id=47fbc4259fba1fb5a9a752bc4a64f977&sr=3&query=recommendation&draft=true (Accessed April 15, 2025)

Windwehr, S. (2025, January 16) '*Systemic Risk Reporting: A System in Crisis?*', EFF. Available at: https://www.eff.org/deeplinks/2025/01/systemic-risk-reporting-system-crisis (Accessed April 15, 2025)

Zuckerberg, M. (2025, January 07) *It's time to get back to our roots around free expression* [Video]. Available at: https://www.facebook.com/zuck/videos/1525382954801931/ (Accessed April 06, 2025)

Legislation

'Commission Implementing Regulation (EU) 2024/2835 of 4.11.2024 laying down templates concerning the transparency reporting obligations of providers of intermediary services and of providers of online platforms under Regulation (EU) 2022/2065 of the European Parliament and of the Council' (2024). *Official Journal L* series. Available at: https://eur-lex.europa.eu/eli/reg_impl/2024/2835/oj/eng

'Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')' (2000) *Official Journal L* 178. Available at: https://eur-lex.europa.eu/eli/dir/2000/31/oj

'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)' (2022) *Official Journal L* 277/1. Available at: https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

Annex I - Overview of the Transparency Database

This Annex presents the database as it is before the Implementing Act from November 2024 enters into force on July 1, 2025, crossing the official Documentation, the DSA legislative text and my experience. It will stay relevant for the most part, except for categories.

Variable name	Type	Values	Explanation	Article DSA
decision_visibility	List (optional)	DECISION_VISIBILITY_CONTENT_REMOVED DECISION_VISIBILITY_CONTENT_DISABLED DECISION_VISIBILITY_CONTENT_DEMOTED DECISION_VISIBILITY_CONTENT_AGE_RESTRICTED DECISION_VISIBILITY_CONTENT_INTERACTION_RESTRICTED DECISION_VISIBILITY_CONTENT_LABELLED DECISION_VISIBILITY_OTHER	When a piece of content's visibility is sanctioned, the nature of the sanction(s) is stated (removed, disabled...)	Article 17(1)(a) - Information on the type of restriction(s) imposed, on the territorial scope, and the duration of the restriction
decision_visibility_other	String (optional)	String	Mandatory precision if "other" is selected above.	
end_date_visibility_restriction	Datetime (optional)	AAAA-MM-JJ hh:mm:ss	No date means a permanent restriction.	
decision_monetary	String (optional)	DECISION_MONETARY_SUSPENSION DECISION_MONETARY_TERMINATION DECISION_MONETARY_OTHER	When a monetary payment is restricted, the nature of the sanction is stated.	Article 17(1)(b) - Information on the type of restriction(s) imposed, on the territorial scope, and the duration of the restriction
decision_monetary_other	String (optional)	String	Mandatory precision if "other" is selected above.	
end_date_monetary_restriction	Datetime (optional)	AAAA-MM-JJ hh:mm:ss	No date means a permanent restriction.	
decision_provision	String (optional)	DECISION_PROVISION_PARTIAL_SUSPENSION DECISION_PROVISION_TOTAL_SUSPENSION DECISION_PROVISION_PARTIAL_TERMINATION DECISION_PROVISION_TOTAL_TERMINATION	When provision is suspended or terminated in whole or in part, the nature of the sanction is stated.	Article 17(1)(c) - Information on the type of restriction(s) imposed, on the territorial scope, and the duration of the restriction
end_date_service_restriction	Datetime (optional)	AAAA-MM-JJ hh:mm:ss	No date means a permanent restriction.	
decision_account	String (optional)	DECISION_ACCOUNT_TERMINATED DECISION_ACCOUNT_SUSPENDED	When an account is suspended or terminated, the nature of the sanction is stated.	Article 17(1)(d) - Information on the type of restriction(s) imposed, on the territorial scope, and the duration of the restriction
end_date_account_restriction	Datetime (optional)	AAAA-MM-JJ hh:mm:ss	No date means a permanent restriction.	

account_type	String (optional)	ACCOUNT_TYPE_BUSINESS ACCOUNT_TYPE_PRIVATE	This specifies the nature of the account connected to the information addressed by the decision (business or personal).	Article 17(3)(b) - Information on the facts and circumstances relied on in taking the decision.
decision_ground	String	DECISION_GROUND_INCOMPATIBLE_CONTENT DECISION_GROUND_ILLEGAL_CONTENT	This specifies whether the decision was taken because the content was illegal (d) or because it was incompatible with the terms and conditions (e)	Article 17(3)(d) Article 17(3)(e) - The legal or contractual grounds relied on in taking the decision
decision_ground_reference_url	String (optional)	String	Where a specific URL to the legal or contractual ground is available, it is encouraged to include it to allow for a quick identification of the ground that was invoked to take the decision.	
illegal_content_legal_ground	String (optional)	String	This specifies the exact legal ground (i.e. the applicable law(s)) that was/were relied upon in taking the decision.	Article 17(3)(d) - The legal or contractual grounds relied on in taking the decision
illegal_content_explanation	String (optional)	String	This explains why the information is considered illegal on the basis of the legal ground indicated.	
incompatible_content_ground	String (optional)	String	This specifies the exact contractual ground (i.e. the relevant section in the applicable terms and conditions) that was relied upon in taking the decision.	Article 17(3)(e) - The legal or contractual grounds relied on in taking the decision
incompatible_content_explanation	String (optional)	String	This specifies why the information is considered incompatible with a specific section in the service's terms and conditions.	
incompatible_content_illegal	String (optional)	String	This specifies whether information that was restricted on the basis of an alleged incompatibility with the terms and conditions was also considered illegal by the online platform.	
category	String	STATEMENT_CATEGORY_ANIMAL_WELFARE STATEMENT_CATEGORY_DATA_PROTECTION_AND_PRIVACY_VIOLATIONS STATEMENT_CATEGORY_ILLEGAL_OR_HARMFUL_SPEECH STATEMENT_CATEGORY_INTELLECTUAL_PROPERTY_INFRINGEMENTS STATEMENT_CATEGORY_NEGATIVE_EFFECTS_ON_CIVIC_DISCOURSE_OR_ELECTIONS STATEMENT_CATEGORY_NON_CONSENSUAL_BEHAVIOUR STATEMENT_CATEGORY_PORNOGRAPHY_OR_SEXUALIZED_CONTENT STATEMENT_CATEGORY_PROTECTION_OF_MINORS STATEMENT_CATEGORY_RISK_FOR_PUBLIC_SECURITY STATEMENT_CATEGORY_SCAMS_AND_FRAUD STATEMENT_CATEGORY_SELF_HARM STATEMENT_CATEGORY_SCOPE_OF_PLATFORM_SERVICE STATEMENT_CATEGORY_UNSAFE_AND_ILLEGAL_PRODUCTS STATEMENT_CATEGORY_VIOLENCE	This high-level classification indicates the main category under which the grounds relied on in a statement of reasons fall.	Article 17(3)(d) Article 17(3)(e) - The legal or contractual grounds relied on in taking the decision

category_addition	List (optional)	STATEMENT_CATEGORY_ANIMAL_WELFARE STATEMENT_CATEGORY_DATA_PROTECTION_AND_PRIVACY_VIOLATIONS STATEMENT_CATEGORY_ILLEGAL_OR_HARMFUL_SPEECH STATEMENT_CATEGORY_INTELLECTUAL_PROPERTY_INFRINGEMENTS STATEMENT_CATEGORY_NEGATIVE_EFFECTS_ON_CIVIC_DISCOURSE_OR_ELECTIONS STATEMENT_CATEGORY_NON_CONSENSUAL_BEHAVIOUR STATEMENT_CATEGORY_PORNOGRAPHY_OR_SEXUALIZED_CONTENT STATEMENT_CATEGORY_PROTECTION_OF_MINORS STATEMENT_CATEGORY_RISK_FOR_PUBLIC_SECURITY STATEMENT_CATEGORY_SCAMS_AND_FRAUD STATEMENT_CATEGORY_SELF_HARM STATEMENT_CATEGORY_SCOPE_OF_PLATFORM_SERVICE STATEMENT_CATEGORY_UNSAFE_AND_ILLEGAL_PRODUCTS STATEMENT_CATEGORY_VIOLENCE vide	This is an optional other category from the same list.	
category_specification	List (optional)	many keywords vide	This is an optional specification from a more granular list of sub-categories.	
category_specification_other	String (optional)	string vide	Mandatory precision if “other” is selected above.	
content_type	List	CONTENT_TYPE_APP CONTENT_TYPE_AUDIO CONTENT_TYPE_IMAGE CONTENT_TYPE_PRODUCT CONTENT_TYPE_SYNTHETIC_MEDIA CONTENT_TYPE_TEXT CONTENT_TYPE_VIDEO CONTENT_TYPE_OTHER	This specifies the type of content that is restricted by the decision to which the statement of reasons relates.	Article 17(1) Article 17(4) Article 17(3)(b) Recital 66 - Submission of clear and specific statements
content_type_other	String (optional)	String	Mandatory precision if “other” is selected above.	
content_language	String (optional)	String	This specifies the language of the piece of content.	Same as above - Submission of clear and specific statements
content_date	Datetime	AAAA-MM-JJ hh:mm:ss	This specifies the date when the platform has started to host the piece of content (date of publication or of account creation).	Same as above - Submission of clear and specific statements

territorial_scope	List	30 codes à deux lettres pour les pays	This specifies the territorial scope of the decision.	Article 17(1)(a)-(d) Information on the type of restriction(s) imposed, on the territorial scope, and the duration of the restriction
application_date	Datetime	AAAA-MM-JJ hh:mm:ss	This specifies the date from which the restriction(s) apply.	Article 17(1)(a)-(d) - Information on the type of restriction(s) imposed, on the territorial scope, and the duration of the restriction
decision_facts	String	String	This specifies the facts and circumstances relied on in taking the decision	Article 17(3)(b) - Facts and circumstances relied on in taking the decision
source_type	String	SOURCE_VOLUNTARY SOURCE_ARTICLE_16 SOURCE_TRUSTED_FLAGGER SOURCE_TYPE_OTHER_NOTIFICATION	This specifies what led to the investigation of the content.	Article 17(3)(b) - Facts and circumstances relied on in taking the decision
source_identity	String (optional)	String	This specifies the identity of the notifier only if that is strictly necessary to identify the illegality of the content.	Article 17(3)(b) - Facts and circumstances relied on in taking the decision
automated_detection	String	Yes No	This attribute indicates whether and to what extent automated means were used to identify the specific information addressed by the decision.	Article 17(3)(c) - Information on the use made of automated means
automated_decision	String	AUTOMATED_DECISION_PARTIALLY AUTOMATED_DECISION_NOT_AUTOMATED AUTOMATED_DECISION_AUTOMATED	This indicates whether and to what extent automated means were used to decide on the infringing nature of the specific information addressed by the decision	Article 17(3)(c) - Information on the use made of automated means
platform_name	String	String	Name of the platform.	
platform_uid	String	String	Unique identifier of the SoR.	
created_at	Datetime	AAAA-MM-JJ hh:mm:ss	Date of creation of the SoR.	

Public Policy Master's Thesis Series

This series presents the Master's theses in Public Policy and in European Affairs of the Sciences Po School of Public Affairs. It aims to promote high-standard research master's theses, relying on interdisciplinary analyses and leading to evidence-based policy recommendations.

Assessing the Digital Services Act's effectiveness in fostering meaningful transparency in social media platforms. The case of Meta.

Lola Pottier

Abstract

Against the opacity and asymmetrical power that characterise digital platforms, the European Union introduced in late 2020 the Digital Services Act (DSA), an ambitious legislation that seeks to promote transparency and accountability in platforms. While the directive's potential benefits and pitfalls have been widely discussed, few studies evaluate how it has translated in practice. This study is a first step towards bridging this gap. Taking Meta as a case study, it assesses to what extent the DSA has been effective in fostering meaningful transparency in social media platforms. It compares Meta's past practices with its current implementation of the DSA, investigating documents produced by Meta, 625M statements of reasons from the Transparency Database and Instagram's user-facing features. The results indicate that the DSA has increased meaningful transparency at Meta, but not equally across individual, aggregate and systemic levels, the three definitional layers of meaningful transparency. While procedural user-facing transparency is quite robust, the DSA has more difficulty fostering understanding of the platform's overall content moderation patterns and impact on society. This can be explained by shortcomings in the DSA itself, but also by the visibility-management strategies that Meta deploys when releasing public information, which hinder its clarity. The thesis concludes by issuing seven recommendations covering two areas: raising the minimum of what is expected of platforms and promoting independent research.

Key words

Transparency, Content Moderation, Digital Services Act (DSA), Meta, Social Media, European Union