Papaevangelou, Charis; Votta, Fabio

Article

Trading nuance for scale? Platform observability and content governance under the DSA

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Papaevangelou, Charis; Votta, Fabio (2025) : Trading nuance for scale? Platform observability and content governance under the DSA, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 14, Iss. 3, pp. 1-31, https://doi.org/10.14763/2025.3.2037

This Version is available at:
https://hdl.handle.net/10419/330355

INTERNET POLICY REVIEW

Volume 14 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Trading nuance for scale? Platform observability and content governance under the DSA

Charis Papaevangelou *University of Amsterdam* c.papaevangelou@uva.nl
Fabio Votta *University of Amsterdam* f.a.votta@uva.nl

DOI: https://doi.org/10.14763/2025.3.2037

Published: 17 September 2025
Received: 11 November 2024 **Accepted:** 20 June 2025

Funding: This work was supported by the Dutch Ministry of Education, Culture and Science under Grant 024.005.017 (Gravitation Research Programme "Public Values in the Algorithmic Society").
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Papaevangelou, C., & Votta, F. (2025). Trading nuance for scale? Platform observability and content governance under the DSA. *Internet Policy Review*, *14*(3). https://doi.org/10.14763/2025.3.2037

Keywords: Digital Services Act (DSA), Platform governance, Platform observability, Transparency, Content moderation

Abstract: The Digital Services Act (DSA) marks a paradigmatic shift in platform governance, introducing mechanisms like the Statement of Reasons (SoRs) database to foster transparency and observability of platforms' content moderation practices. This study investigates the DSA Transparency Database as a regulatory mechanism for enabling observability, focusing on the automation and territorial application of content moderation across the EU/EEA. By analysing 439 million SoRs from eight Very Large Online Platforms (VLOPs), we find that the vast majority of content moderation decisions are enforced automatically and uniformly across the EU/EEA. We also identify significant discrepancies in content moderation strategies across VLOPs, with TikTok, YouTube and X exhibiting the most distinct practices, which are further analysed in the paper. Our findings reveal a strong correlation between automation and the speed of content moderation, automation and the territorial scope of decisions. We also highlight several limitations of the database, notably the lack of language-specific data and inconsistencies in how SoRs are reported by VLOPs. We conclude that despite such shortcomings, the DSA and its Transparency Database may enable a wider constellation of stakeholders to participate in platform governance, paving the way for more meaningful platform observability.

Introduction

The Digital Services Act (DSA; Regulation (EU) 2022/2065, 2022) has been described as a "transparency machine" (Zornetta, 2024) with the potential of creating a "global transparency regime" (Helberger & Samuelson, 2024). The DSA, among others, requires platforms[1]to provide a justification for their content decisions (e.g., grounds for content removal or restriction) in the form of a Statement of Reasons (SoRs; DSA Art. 17(1)) to shed light on the often-opaque decision-making processes of their content moderation systems. These SoRs are then automatically uploaded and aggregated to a publicly accessible online database operated by the European Commission (DSA Art. 24(5)) putting in place a system of "automated transparency" (Kaushal et al., 2024).

The early impact of the DSA is already apparent. For example, researchers have uncovered inconsistencies in the self-reported practices of X, which claims to rely exclusively on manual moderation in its SoRs submissions but acknowledges using automated tools in its transparency reports (Dergacheva et al., 2023; Drolsbach & Pröllochs, 2023). Such cases highlight the latent regulatory potential of the DSA and demonstrate how transparency can enable *platform observability* (Rieder & Hofmann, 2020), a dynamic and analytical governance mechanism that facilitates a deeper understanding of platforms' governance practices (Leerssen, 2024).

In this paper, we approach the Transparency Database as an instrument for platform observability, which builds upon transparency, to conduct an empirical case study on the kinds of insights that can be inferred from the Transparency Database. We seek to understand the extent to which the SoR database fosters observability, which we recognise as a governance mechanism that is necessary, in addition to transparency, to ensure platform accountability and deepen our understanding of how platforms operate and exercise their power over our digital public spheres (van Dijck et al., 2018). Our study, specifically, focuses on exploring the use of automation in content moderation and the differences in the territorial scope of content moderation practices, as well as the implications thereof for the DSA and platform governance.

We conducted our analysis with two leading questions in mind: how do content governance decisions vary among platforms and member-states in the EU, and in what ways does the use of automation in moderation differ across platforms, the

1. The DSA covers all intermediary platforms, including e-commerce and marketplaces (e.g., Zalando, App Store), ride-sharing apps (e.g., Uber), and others.

EU and its member-states. We concentrated our analysis on the moderation practices of eight digital Very Large Online Platforms (VLOPs), namely X, Facebook, Instagram, YouTube, TikTok, Snapchat, Pinterest and LinkedIn, most of which have been foundational for our modern digital public spheres and the broader digital platform ecosystem.

We used R, a programming language that is widely used for data analysis (R Core Team, 2021), to retrieve 439 million SoRs from a period of four months (25 September 2023 to 25 January 2024). We conclude that the DSA enables a variety of actors (Helberger et al., 2018), including researchers and members of civil society, to make use of technically-sophisticated regulatory mechanisms, such as the Transparency Database, supplementing the more traditional governance mechanism of transparency. In that sense, the Transparency Database, despite its shortcomings, marks a crucial turning point in governing digital platforms as it paves the way for observing platforms' behaviour and content moderation practices in a way that was unattainable before. Our paper contributes to a growing body of literature on platform governance in regulated digital environments, especially in the EU. In doing so, it also offers a critical perspective on the DSA's regulatory and governance ambitions.

In summary, we discerned three key findings. First, we found that 99% of all moderation decisions were applied uniformly across the EU/EEA, exhibiting a tendency for uniform application of content moderation strategies. Second, we identified variations in how VLOPs enforce content moderation, particularly in terms of regional differences, with only three, namely TikTok, Youtube and X, reporting conducting territorially specific content moderation. Third, we observed that the use of automation in content moderation correlates with faster enforcement timelines and uniform application, whereas manual moderation often means longer delays and more territorially specific application. This last finding also raises questions about the interaction of EU and national legal frameworks and the VLOPs' capacity to handle "low-resource languages" (Nicholas & Bhatia, 2023). Last, we consider some technical shortcomings of the database and regulatory blindspots of the DSA such as the lack of an obligation to report the language of the content that was moderated. The next sections are structured as follows: first we situate our paper theoretically through a literature review of works relevant to content moderation, platform governance and regulation in order to build our case for platform observability; second, we expound on our methodological approach; third, we present our findings in detail; finally, we discuss the implications of our findings for the DSA and platform governance.

Conceptual framework: Content moderation, platform governance and the need for platform observability

Our conceptual framework is developed in three steps: first, we do a brief literature review of scholarship on content moderation as an instrumental part of platform governance and, subsequently, platforms' power; second, continuing our literature review, we trace the rise of automation in content moderation, along with its structural limitations; and third, we advance the need to adopt platform observability as a critical lens through which to study platform governance and as a fitting regulatory approach, as we later show through our case study on the DSA.

Content moderation and platform governance

In recent years, a rich body of research from critical media to legal scholarship has extensively explored the politics and mechanisms of content moderation and its implications for democracy. Without attempting an exhaustive discussion of the literature, we draw upon scholarship on platform governance which has illuminated how platforms navigate the balance between self-regulation and state oversight, highlighting the informal and formal mechanisms that underpin platform governance (Gorwa, 2019b, 2019a; Papaevangelou, 2021). Gorwa has also, crucially, underscored the role of power dynamics between platform governance stakeholders in influencing content moderation, taking stock of the political reality within which platform governance is inscribed (Gorwa, 2024). In a similar vein, Griffin (2023) has theorised about the politics behind content moderation decisions, demonstrating how platforms balance legal, public and commercial interests in their governance strategies. This strand of scholarship has, therefore, showcased the inherently political process of moderating content that is often obscured by platforms' sophisticatedly opaque systems, which include extractive processes of invisible labour (Roberts, 2019) and which are veiled behind discourses of neutrality (Gillespie, 2010).

Subsequently, content moderation is closely tied to the platforms' political power, that is, the power to shape the norms, rules and conditions under which information circulates online and, by extension, influence the structures of our digital public spheres and the capacity of citizens to form political opinions (Gillespie, 2018; Helberger, 2020). Content moderation, thus, emerges not simply as a technical

function but a central mechanism of platform firms (Grimmelmann, 2017), governing the socioeconomic, cultural and political interactions of end-users, complementors and other relevant actors that convene in the multi-sided markets that large platforms constitute (see Poell et al., 2021 Chapter 4). From this perspective, social media platforms rely heavily on content moderation to maintain their advertising-driven business models. The ability to regulate, filter and organise user-generated content enables platforms to create an environment conducive to their revenue goals, whereby the goal is to maximise profits through the increase of user engagement, while minimising harmful or controversial content that could alienate end-users or advertisers (Griffin, 2022; Jimenez-Duran, 2022).

Automation in content moderation and its structural limitations

Artificial Intelligence (AI) and automated decision-making processes (Bloch-Wehba, 2020; Gillespie, 2020; Gorwa et al., 2020) have been crucial in dealing with the massive volume of user-generated content uploaded every instant on social media platforms, while reducing the cost of content moderation (e.g., instead of hiring human reviewers). Typically, AI refers to systems that use machine learning models to identify, classify, or predict patterns in data. Combined with automation and automated-decision making, which typically refer to programmed processes that operate without human intervention (at least not necessarily visible labour), these sociotechnical systems have given rise to the model of "algorithmic commercial content moderation" (Gorwa et al., 2020, p. 3). The latter refers to "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome" (Gorwa et al., 2020, p. 3).

The level of automation varies depending on the type of content it is deployed against and the specific legal framework in place. For instance, the identification of terrorist or copyrighted pieces of content is a predominantly automated process based on hash-matching, whereas toxic or hateful content might involve a combination of automation and human reviewing (Gorwa et al., 2020, p. 7). Algorithmic moderation is also veiled with a neutrality discourse, obscuring the aforementioned political-economic dimension of platform governance and emerges as "a shared imaginary of technological solutionism" (Udupa et al., 2023, p. 1). Automation and technology, in this context, are presented as providers of efficient, accurate and swift solutions to issues relevant to our democracy (cf. Poell et al., 2021 Chapter 7). As a result, relevant scholarship has noted an increasing reliance on and embracing of automation to "sanitise" online spaces (Griffin, 2022), often at

the expense of benign, marginalised, or non-conforming discourses (Are, 2022).

In that sense, automation in moderation, and the discursive embrace thereof, not only obfuscates its shortcomings but, more broadly, seeks to depoliticise platform governance and blur the messy reality of its political economy (Gorwa et al., 2020, p. 13). Moreover, this dynamic tends to obscure the critical-yet-undervalued human labour activities like data annotation and content reviewing, which are essential for training the models undergirding the automated processes of content moderation and, largely, for the political economy of content moderation (Posada, 2022; Roberts, 2019). Such labour, often outsourced in the *majority world,* is at the core of platforms' content moderation industrial approach (Caplan, 2018), that is a massive automated, systematised and standardised system of moderation designed for scalability and cross-jurisdictional application. Last, the inadequacy of automation in content moderation becomes even more pronounced when engaging with "low-resource languages", that is, languages that have a scarce digital footprint and, thus, have not been incorporated in the training of these models (Nicholas & Bhatia, 2023)

The limits of transparency and the turn to platform observability

These structural problems and limitations of industrial automated content moderation highlight the inadequacy of current governance mechanisms that are rooted in techno-solutionism. It is no surprise, then, that the response, primarily by platforms but also some policymakers and experts, was to ask for (more) transparency to unravel the "black box" (Pasquale, 2016). Indeed, faced with public outcry, major social media platforms put various transparency mechanisms in place, ranging from self-reported content moderation transparency reports to establishing specialised teams–the industry term is 'Trust and Safety'–dealing with content moderation (Gorwa & Ash, 2019). These developments led platforms to "networked platform governance" structures that involve a combination of automated means, dedicated workforce and the implication of external stakeholders to govern their digital spaces and, purportedly, foster trust in involved stakeholders (Caplan, 2023). In fact, in this novel configuration, companies over-emphasised the role of transparency and of "openness and access […] to establish trustworthiness to external actors (Caplan, 2023, p. 3462).

However, as critical platform scholars have noted, transparency can only go so far as a platform governance mechanism due to the inherent sociotechnical complexities of digital platforms (Ananny & Crawford, 2016), pushing for the adoption of

platform observability as a more suitable governance mechanism (Rieder & Hofmann, 2020; Leerssen, 2024). Indeed, transparency is often considered and critiqued as a static (Leerssen, 2024) form of "visibility management", that is a highly curated and mediated process of disclosure of data, information and insights (Flyverbom, 2019, p. 18) by platforms to external stakeholders. But the underlying premise of transparency as a prerequisite component to "create the knowledge required to govern and hold systems accountable" (Ananny & Crawford, 2016, p. 975) holds. Hence, scholars studying platform governance do not argue for abandoning transparency[2]but rather complementing it with observability and going even further.

Drawing on recent works on platform observability (Rieder & Hofmann, 2020; Ferrari et al., 2023, Leerssen, 2024), we conceptualise it as a continuous and process-oriented approach that acknowledges the complexities of observing and understanding digital platforms. Unlike transparency, observability is an active, pragmatic and explicitly subjective practice, decentering our focus on "the algorithm" to encompass more facets of the processes that constitute platform infrastructures and their automated-decision making systems (Leerssen, 2024, pp. 7-8). Rieder and Hofmann, specifically, set three principles as foundational for observability: observability in relation to, or in favour of the public interest (normative principle); observability as a continuous and dynamic process (sociotechnical principle); observability as a catalyst for collaborative forms of platform governance (governance principle). Additionally, the authors connect the concept directly to regulation, arguing for a dual approach involving "regulating for observability" and making observability part of our regulatory strategies for platform firms (Rieder and Hofmann, 2020, p. 22).

We further understand observability as an attempt to reverse neoliberal framings of transparency as objective, which sought to depoliticise the notion and disentangle it from its political-economic, social and cultural power structures (Birchall, 2021). Therefore, we conceptualise observability not simply as constrained to the disclosure of information but as *the regulatory capacity to continuously monitor and scrutinise platform operations,* while considering the emergent nature of platform behavior, influenced by user interactions, sociopolitical and economic strategies of platform owners and platforms' automated-decision making algorithms. Consequently, observability does not simply enable a technical understanding of platform technologies but sets in place the architecture for a network of actors (Ca-

2. For a comprehensive comparison of transparency and observability, including of their metaphors, see Leerssen, 2024.

plan, 2023), beyond traditional and often-opaque regulatory processes like auditing (Terzis et al., 2024) to actively and cooperatively participate in digital governance (Helberger et al., 2018; Keller & Leerssen, 2020). In that sense, observability complements and goes further than transparency to empower governance actors to act in an informed manner, reduce information asymmetries and increase trust among stakeholders and, generally, improve the conditions necessary for accountability (Birchall, 2021, p. 6; Gorwa et al., 2020, p. 11). Taking the sociotechnical aspect of platform observability "as a regulatory program" (Leerssen, 2024, p. 9) a step further, we hold that it should also include more opaque activities that exist in platforms' value chains and which enable their operations, ranging from labour relations (e.g., real-time information about human moderators employed or contracted) to environmental implications (e.g., real-time insights into computing power used along with its environmental footprint). If we are to leave the "snapshot logic" (Rieder & Hofmann, 2020, p. 7) behind for a more comprehensive and holistic regulatory framework, then we must consider the labyrinthine, interrelated complexities of platforms' political economy.

Building on Rieder and Hofmann's conceptual foundation, Ferrari and colleagues (2023) propose a complementary framework for the governance of generative AI systems, which introduces three "oversight conditions" (p. 2) for observability: industrial observability (the capacity to scrutinise the interlinked, material and political-economic layers of generative AI models and systems from a macro-level), public inspectability (the need for deep regulatory inspection to all layers constituting such systems) and technical modifiability (the capacity to make fundamental changes to these systems to ensure compliance with oversight mechanisms). The similarities are evident in the prominence given by both frameworks to the public interest and the role of public regulation, but what stands out is the "technical modifiability" condition as it seems to indicate and advocate for further action – beyond regulatory – to be taken as a response to the insights we may derive from observing platform and AI systems. Arguably, this last condition shows that observability too must be further expanded when dealing with tech companies and their complex value chains.

The Digital Services Act and regulated platform observability

The EU's DSA (Regulation (EU) 2022/2065, 2022), voted in 2022 and enacted in full force in 2024, imposes due diligence obligations on platforms to ensure a safe

transparency and predictable online ecosystem and the protection of users' fundamental rights online (Art. 1 DSA). The DSA, moreover, foresees financial penalties of up to 6% of the company's global turnover in case of regulatory infringements (DSA Art. 52 (3)). It aims at enhancing transparency in the decision-making processes of content moderation and, largely, regulating how platforms moderate their digital spaces. While many due diligence obligations apply to hosting services and platforms in general (Articles 16-28), the DSA introduces additional obligations for Very Large Online Platforms (VLOPs), particularly in relation to systemic risk mitigation and oversight of automated content moderation processes (Articles 33-43). In that sense, the DSA adopts a systemic approach to platform regulation, whereby it focuses "mostly on the process and design rather than the content itself" (Husovec & Roche Laguna, 2022, p. 1) to ensure that platforms' content governance systems mitigate their contribution to "systemic risks" like the spread of illegal content or threats to the exercise of fundamental rights (DSA Art. 34 (1)). In this context, VLOPs are expected to enact risk mitigation mechanisms, including regular risk assessments and third-party audits (Terzis et al., 2024).

As noted in the introduction, the DSA has been characterised as a "transparency machine" (Zornetta, 2024) with the potential to set in place a "global transparency regime" vis-à-vis platform governance (Helberger & Samuelson, 2024). Indeed, one of the key aspects of the DSA, which is also at the heart of this paper, is its capacity to compel platforms, especially VLOPs, to make their (automated) decision-making processes more transparent, legible, hereby increasing the opportunities of relevant stakeholders (i.e., civil society organisations, researchers, regulatory authorities) to contributing to holding platforms accountable. Specifically, provisions like the obligation for platforms to issue a Statement of Reasons (SoR) to afflicted users for each content moderation decision (DSA Art. 17 (1)) are important steps to reinforcing transparency and, thus, enabling us to understand platforms' processes better (Leerssen, 2024). Briefly, these SoRs must include key details, such as the nature of the decision (e.g., content removal, restriction, demotion, etc.), the supporting facts and pertinent legal or contractual grounds for the decision, the reporting of use and level of automation in identifying content and enforcing decisions, and the inclusion of "user-friendly" information about available redress mechanisms (Art. 17(3)(a)–(f)).

Yet, the paradigmatic change brought about by the DSA in platform governance is that it introduces principles of platform observability. To illustrate, Art. 17(3)(c) DSA requires that the SoRs include information regarding the use of automated means in the content governance decision, including detection and enforcement

(Art. 17(3)(c) DSA). These SoRs are then compiled and made public in a "machine-readable" database–the DSA Transparency Database which is at the core of this paper's focus–that is managed by the Commission (Art. 24(5) DSA). Other relevant–but beyond this paper's scope–provisions of platform observability are those that concern access to data for vetted researchers and other interested actors to study topics related to (potential) systemic risks stemming from platforms (DSA Art. 40 (1)(12) DSA). As Leerssen (2024, p. 28) notes, such data and information disclosure rules in the DSA reflect a (platform) observability approach. Effectively, these provisions showcase how the DSA goes beyond the typical disclosure of information and sets the conditions for ongoing, collaborative and situated scrutiny of platforms.

On the website of the DSA Transparency Database, we read how it enables "transparency and scrutiny" and "[monitoring] of [the spread] of illegal and harmful content online" through access to machine-readable data in "almost real-time". The machine-readable and real-time aspects are crucial as they allow for a more technically sophisticated analysis of the SoRs and a necessary level of consistency across all different platforms, which was lacking in previous regulatory attempts of the Commission like the voluntary Code of Practice on Disinformation. The Transparency Database also provides insights into the territorial aspect of content moderation processes, that is, what differences there exist in how platforms enact their content policies across the EU and its member states. These insights are important to assess the challenges of having a common legal framework under the DSA for content governance in the EU, which itself does not ensure legal harmonisation when it comes to "what content or behaviour counts as illegal" (Husovec & Laguna, 2022, p. 11) as that capacity largely still rests with member-states' legal frameworks, except in certain areas like hate speech, where illegality is defined at the EU level. In fact, as Husovec and Laguna point out, platforms, as the "delegated enforcers" of the DSA (Husovec, 2024), are expected to act as enforcers of the law, always prioritising a decision that makes them to comply with the law's obligations, even if that means erring on the side of over-censoring to avoid legal repercussions (Keller, 2024).

Further, emerging empirical research has begun to analyse the SoR Database, which is proving to be a methodologically challenging site for accountability. Drolsbach and Pröllochs (2024) conducted a large-scale quantitative study of 156 million SoRs across the same major social networking platforms as ours over the first two months of the database's launch. Their analysis reveals significant inconsistencies in how platforms interpret and implement their reporting obligations.

For instance, moderation frequency varied drastically, with TikTok submitting the most (predominantly automated) SoRs than the rest of the cohort by far – more than 100 million SoRs and over 350 times more per user than X (Drolsbach & Pröllochs, 2024, p. 939). Additionally, they found how platforms frequently labeled content using categories not-predefined by the DSA (i.e., "Other"), as well as how "incompatibility" with platforms' rules dominated over obligations to tackle illegal content (Drolsbach & Pröllochs, 2024, p. 940). Similarly, Trujillo and colleagues (2025) conducted a quantitative study of the first 100 days of the Database's operation, analysing 353 million SoRs from the same eight major platforms. Crucially, the authors, same as we did, cross-compared information data using platforms' transparency reports, allowing them to identify various inconsistencies such as how X submitted only 466 thousand SoRs to the Commission while reporting over 2 million moderation actions for the same period (Trujillo et al., 2025, p. 15). Both studies highlight that the data are self-reported, which leads to significant inconsistencies as platforms still wield significant discretion in implementing their regulatory obligations.

Last, another significant work comes from Kaushal and colleagues (2024) who analysed 131 million SoRs from 15 platforms from November 2023, combining their quantitative study with a legal analysis of the DSA through the lens of the useful concept of "automated transparency". While the insights and findings largely corroborate those of the other empirical studies, the authors also foregrounded the deficiency in consistently reporting the language of the content that was moderated (Kaushal, 2024, p. 1128), as the DSA does not oblige VLOPs to report the language of the affected content. Therefore, the Transparency Database contains minimal information on that front, which is a significant limitation when it comes to understanding how content moderation is not only affected by the legal differences across member-states but also by the shortcomings or biases of platforms' content moderation systems as regards the employment of human reviewers who can understand the various languages spoken in the EU and its cultural diversity. All in all, these studies also show the promise that the Transparency Database – and broadly the DSA – holds for platform observability. It is precisely thanks to this Database that inconsistencies like that of X could be foregrounded, which may now be used by the Commission in its proceedings against the platform (European Commission, 2023). Yet, we have to be cautious as the Database can also be used for performative compliance, whereby platforms may "strategize their use of the database as a means to show their compliance" (Kaushal et al., 2024, p. 1130). Building also, then, on this emerging literature and combined with our conceptual framework, we approach the Transparency Database as a *socio-technical infrastruc-*

ture of observability, which produces knowledge about platforms' infrastructures, through specific conditions which are unfortunately greatly shaped by platforms themselves, but which nevertheless makes platform governance observable and contestable. In the following sections, we present our empirical analysis of the Transparency Database and discuss our findings.

Research objective and methods

The primary aim of our study is to determine the extent to which insights on automated content moderation can be derived from the Transparency Database of the DSA, particularly concerning language representation, consistency in applying moderation policies across the EU and discrepancies among platforms. As a result, we conducted our analysis with two leading questions in mind: how do content governance decisions vary among platforms and member-states in the EU, and in what ways does the use of automation in moderation differ across platforms, the EU and its member-states.

From an observability point of view, these questions are crucial because they touch on the core challenge of making platform governance not only transparent but also of making the behaviour of platforms observable and open to scrutiny to external stakeholders, including researchers. Subsequently, the parameters we looked for in our data retrieval were: name of platform, means of detection, means of decision, territorial scope, language of content, dates of content and decision. We also studied the transparency reports self-reported by the same VLOPs for data on Average Monthly Active Recipients (AMARs)and human moderators, spanning the EU, its member-states, and official languages (Tables 3 & 6). By integrating these data sources, we wished to obtain a more accurate depiction of the relationship between automation, country and language to explore regional or national disparities in moderation.

As mentioned in the introduction, we were exclusively interested in studying digital platforms that have effectively become the infrastructure of our digital public spheres and that comprise the digital platform ecosystem in Europe, and broadly the West. Our study, thus, concentrated on eight digital VLOPs: X, Facebook, Instagram, YouTube, TikTok, Snapchat, Pinterest and LinkedIn, encompassing different platform types (e.g., social media, video-sharing, microblogging, image-based and professional networks). Moreover, as we elaborate in the next section, following our initial analysis we found that only three–X, YouTube and TikTok–exhibited sufficient variation in their content moderation practices, especially with respect to

territorial differentiation and the use of automation which were key to our research, resulting in a more in-depth subsequent analysis into these three platforms.

Using R to retrieve their SoRs submissions over four months (25 September 2023 to 25 January 2024[3]), we created a data set of 439 million SoRs (Table 1). We also normalised the SoRs per 1,000 AMARs to account for user base variances across countries, particularly focusing on whether moderation is disproportionately applied to certain areas. For instance, in Table 4 we see that EEA ranks 5th concerning YouTube's content moderation despite having the largest volume of SoRs (18M); that is because they produce approximately 40 SoRs per 1,000 AMARs in contrast to, for example, the Netherlands where they produce almost 80 SoRs per 1,000 AMARs. In that sense, our normalisation also underscores the intensity of moderation relative to the number of SoRs in specific member-states or regions (e.g., EEA/EU). In the next section, we discuss our main findings concerning the implications of the DSA in relation to platform governance, algorithmic content moderation and platform observability.

TABLE 1: Statement of Reasons per platform

PLATFORM	STATEMENTS OF REASONS	AVERAGE MONTHLY ACTIVE RECIPIENTS (AMARS)
FACEBOOK	95,787,720	259,000,000
INSTAGRAM	11,218,806	259,000,000
LINKEDIN	48,666	45,200,000
PINTEREST	72,661,464	124,000,000
SNAPCHAT	1,398,507	101,973,520
TIKTOK	235,107,438	125,000,000
X	608,168	126,120,951
YOUTUBE	22,102,346	416,600,000

Limitations of our study

It must also be stressed that we faced significant limitations in our analysis. Firstly, a major drawback of the database vis-à-vis platform observability, thus, is the lack of contextual information about the decisions. Instead, we are left with a reduced understanding of the reality that is unfolding on and through social media plat-

3. The DSA Transparency Database was launched on 25 September 2023.

forms. To illustrate, although platforms must indicate on what grounds they took a certain decision (e.g., which term of service did a piece of content violate), this is done in a quasi-machinist way that offers little to no sensible information. Leerssen notes, for instance, that "[a] more ambitious approach would have included URLs or other unique identifiers, where possible, to link decisions to specific content items" (2024, p. 21).

Secondly, the database is in constant development, meaning that the operators at the Commission are still refining its features. For instance, there were incidents of inconsistencies in our data-gathering process due to unannounced technical changes in the database concerning the method of data compilation, forcing us to recommence from scratch our scraping. Thirdly, the SoRs and the transparency reports are self-reported by VLOPs, which, as demonstrated, are not entirely reliable in providing accurate information.

In sum, these limitations did not allow us to conduct a more fine-grained analysis of the available data to discern, for instance, which types of content were more likely to have a specific territorial scope. Likewise, we could not derive explicit insights from the relation of automation and language on a specific content level because none of the VLOPs in our sample reported the language of the content that was moderated[4]. In the future, such limitations could be partially remedied by securing access to finer data, for instance through Article 40 of the DSA, which grants access rights to data for vetted researchers (Leerssen, 2024).

Findings

Our analysis focuses on eight digital VLOPs, for which we retrieved and analysed 439M SoRs from the DSA Transparency Database over a period of four months. Our analysis revealed three key patterns: first, content moderation decisions are overwhelmingly applied uniformly across the EEA or EU, with 99% of SoRs exhibiting no territorial differentiation; second, only three platforms of our corpus (YouTube, TikTok and X) report engaging in territorially specific moderation practices; and third, the degree of automation in detecting and enforcing moderation decisions seems to significantly affect the timelines of enforcement, with manual reviews correlating with longer delays. Regarding our first finding, we must note that the remaining 1% still comprises millions of SoRs, indicating that territorially specific moderation, while limited, is not negligible. In the following paragraphs, we elabo-

4. We discovered that only a few platforms outside the social media ecosystem (e.g., Google Shopping) reported language-specific data.

rate on our findings relating to these two central dimensions of platform governance: the territorial scope of content moderation decisions and the automation of detection and enforcement processes.

Territorial aspect of content moderation

One of the most striking insights from our analysis is the territorial distribution of SoRs across the European Economic Area (EEA), EU and individual member-states. The top two territorial scopes concern the EEA (with and without Iceland and Norway; Table 2), accounting for 50.40% (or 221M SoRs) and 48.41% (or 213M SoRs) of all SoRs, while the third one concerns the EU, accounting for 0.32% of our corpus (or 1M of SoRs). Put simply, almost all moderation decisions (99.13% or 435M out of the 439M SoRs of our corpus) made by VLOPs in our sample were applied uniformly across the entire EEA and EU. As regards specific member states, we found that Germany leads in terms of country-specific SoRs, with 588K SoRs (0.13%), followed by France (432K, 0.10%) and Italy (323K, 0.07%); this is to be expected given that these countries host the largest populations in the EU.

To infer better country- and language-specific insights, we decided to eliminate platforms whose content moderation decisions were applied uniformly across the EU or EEA. This approach narrowed our focus to the remaining 0.87% (or approximately 4M SoRs), where territorial variation was evident. To illustrate, Facebook's (in blue colour; Figure 1) content moderation decisions are almost exclusively applied on an EU and EEA level, providing no meaningful divergence for analysis at the member-state level.

TABLE 2: Top 10 territorial scope of content moderation decisions in the EU/EEA

RANK	TERRITORIAL SCOPE	SORS	% SORS
1	EEA (NO ICELAND)[5]	221M	50.40%
2	EEA	213M	48.41%
3	EUROPEAN UNION	1M	0.32%
4	GERMANY	588K	0.13%
5	FRANCE	432K	0.10%
6	ITALY	323K	0.07%
7	EEA (NO ICELAND OR NORWAY)	197K	0.04%
8	POLAND	185K	0.04%
9	SPAIN	169K	0.04%
10	FINLAND, HUNGARY, LIECHTENSTEIN, LITHUANIA, NORWAY, POLAND, SLOVENIA	129K	0.03%

Therefore, we focused on YouTube, X and TikTok, as they were the only platforms whose data showed some variation in content moderation decisions across different member-states (Table 4). TikTok applied most of its content moderation decisions uniformly across the EEA level. Simultaneously, the platform dominated our corpus showcasing the dominance of the video-sharing platform in the EU market (235M SoRs, 53,54%). As such, TikTok demonstrates a preference for an EEA-wide content moderation, with minimal variation between individual countries. When it engages in country-specific moderation (e.g., in the Netherlands or France), the volume is significantly lower than its regional enforcement, which is in line with a more industrialised and streamlined content moderation strategy. YouTube's data indicates a more nuanced approach to territorial content moderation than TikTok, which is also demonstrated by the geographic dispersion of its decisions across areas in our scatterplot (Figure 1). YouTube is, thus, more likely to apply a territorialised approach to moderating content in individual member-states, the Netherlands, France and Italy in particular, reflecting a greater variety in its overall strategy, in spite of applying most of its decisions on an EEA level. Lastly, X reports to be

5. THE TERRITORIAL SCOPES IN TABLE 2 REFLECT SELF-REPORTED LABELS IN THE DSA TRANSPARENCY DATABASE. DUE TO INCONSISTENT LABELLING ACROSS PLATFORMS AND LACK OF PROOF-CHECKING MECHANISMS BY THE COMMISSION, OVERLAPPING CATEGORIES SUCH AS 'EEA' AND 'EEA (NO ICELAND)' MAY REFER TO THE SAME GEOGRAPHIC SCOPES, AFFECTING THE ACCURACY OF DISTINCT TERRITORIAL CATEGORIES' NUMBERS.

implementing its content moderation decisions in a highly territorialised manner, with the Netherlands ranking again the highest among individual member-states.

Automation in content moderation

We then proceeded in visualising the data in two scatterplots to represent the degree of automation and its impact on the speed of enforcement of content moderation decisions across different territories (Figures 1 & 2). For clarity, in the first scatterplot (Figure 1), the horizontal axis represents the degree of automated detection, ranging from manual detection (on the left) to fully automated detection (on the right), while the vertical axis represents the volume of SoRs per 1,000 AMARs, with the sizes of the circles representing the volume of SoRs in a region. In the second scatterplot (Figure 2), the horizontal axis represents the same as the first scatterplot, while the vertical axis represents the median day of the application of content moderation decisions.

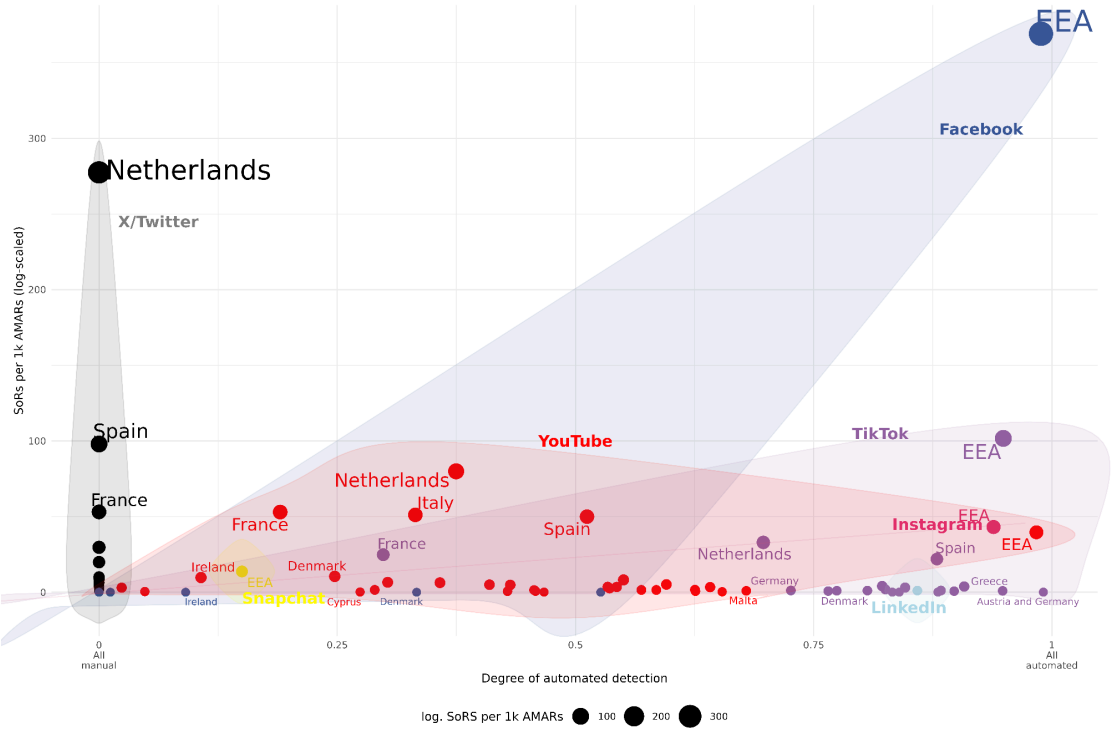

Figure 1 - Statement of Reasons concerning automatically or manually detected content per 1,000 Monthly Active Users

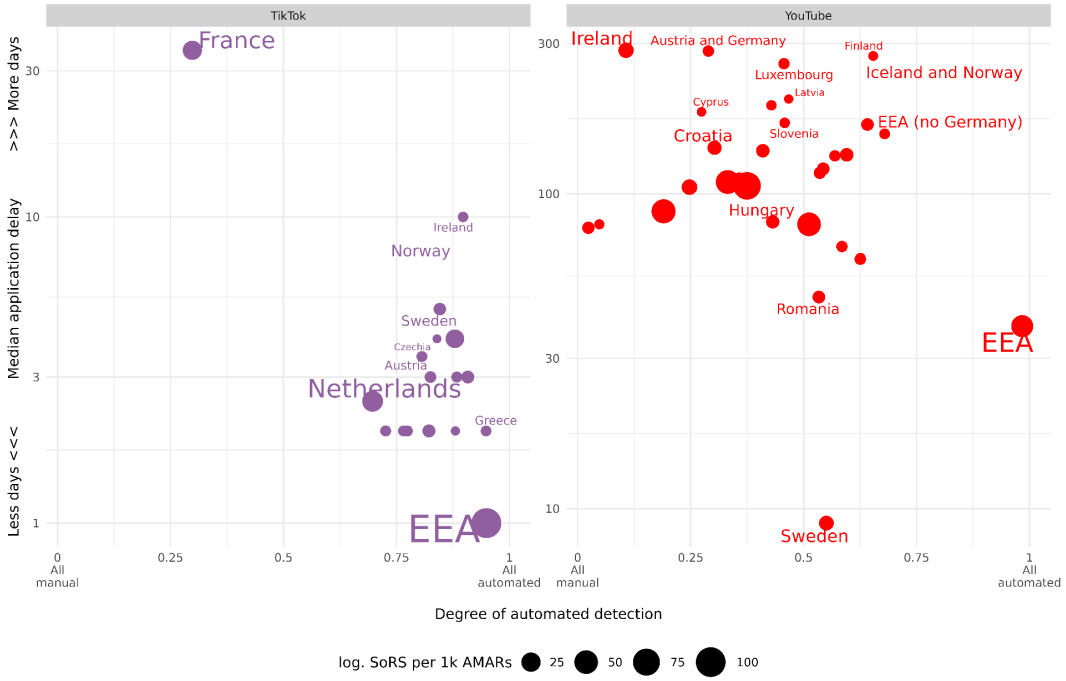

Figure 2 - Scatterplot showing the relation between time and automation of detection with the territorial scope of content moderation decisions for TikTok (left-hand side) and YouTube (right-hand side)

Our findings here also reveal significant variations. TikTok shows a high reliance on automated detection, particularly when moderating content on an EEA-level but also in some countries like Spain and the Netherlands (Figure 1). Notably, as the case of France shows in both scatterplots, when content is detected manually it is also more likely to be dealt manually, taking significant time for a decision to be applied. For example, the median day of enforcement for TikTok in France in our sample was more than 30 days (Figure 2). Put simply, the shorter enforcement delays in these countries suggest that TikTok uses automation to address most of the content.

YouTube seems to employ more of a hybrid approach, predominantly relying on automated detection methods supplemented with manual review. Generally, the same pattern as TikTok was also observed in this case, with manual detection demonstrating a correlation with manual enforcement. What is more, with a closer look in our data, we discerned that YouTube's median reaction time, including automatically detected content, can extend in occasions over 100 days (Figure 2), with delays increasing dramatically when manual review is involved and when decisions concern specific territorial scope (e.g., France).

In contrast, X stands out as an outlier, reportedly relying exclusively on manual moderation. The reason why X's data was not visualised as a time scatterplot in our analysis is that the platform indicates moderating all content manually and on the same day as detected (Figure 1). As mentioned earlier, this observation contradicts its public transparency reports which mention the use of automated means[6], making X's data unreliable. This discrepancy has been also corroborated by other studies (Dergacheva et al., 2023; Drolsbach & Pröllochs, 2023; Kaushal et al., 2024).

Finally, our tables (3 & 6) containing data on human reviewers indicate that languages such as French and Dutch receive more attention from human moderators across platforms, especially for YouTube and TikTok. YouTube, for instance, allocates 176 French-speaking moderators and 24 Dutch-speaking moderators, while TikTok allocates 687 French-speaking moderators and 167 Dutch-speaking moderators, the latter being notably high compared to other platforms. Again, given that these are self-reported figures using non-standardised methodologies, any cross-platform comparison remains inherently limited. This could suggest that better language coverage allows platforms to manually moderate more content and, thus, provide–arguably–more accurate decisions that, as shown in Figure 2, take more time than automated decisions.

Having said that, the number of human moderators is not correlated to the use of automation in content moderation. As Klonick's (2018) analysis of Facebook's content moderation system has shown, these platforms have different tiers to address content moderation issues according to their perceived severity and importance (see also Caplan & Gillespie, 2020). In other words, it might very well be that TikTok still predominantly relies on automated means to moderate content but employs–the most–human reviewers to deal with more sophisticated issues (Table 3). Therefore, a key limitation in understanding the operational capacities of platforms across EU member-states lies in the heterogeneity of reporting metrics regarding human moderation staff.

6. Available here: https://transparency.x.com/dsa-transparency-report.html.

TABLE 3: Human moderators employed by platforms based on their transparency reports

PLATFORM	PERIOD	AMARS	MODERATORS	AMARS/MODERATOR
FACEBOOK	01/04/2023-30/09/2023	259.000.000	1.362	190.161
INSTAGRAM	01/04/2023-30/09/2023	259.000.000	1.362	190.161
YOUTUBE	01/01/2023-30/06/2023	416.600.000	1.974	211.043
LINKEDIN	01/01/2023-30/06/2023	45.200.000	146	309.589
SNAPCHAT	01/02/2023-30/07/2023	101.973.520	1.545	66.002
PINTEREST	01/03/2023-30/06/2023	124.000.000	1.963	63.168
X (TWITTER)	01/04/2023-30/10/2023	126.120.951	2.496	50.529
TIKTOK	01/04/2023-30/09/2023	125.000.000	5.827	21.451

Discussion

Content and platform governance in the EU

Our analysis showcases how the DSA and, the Transparency Database in particular, is a step toward fostering a dynamic way of studying platform governance and, thus, in enabling platform observability. Our findings foreground the inherent tension in the pursuit of faster and more accurate content moderation decisions, especially via the use of automation. They also demonstrate how calls for faster reaction times inherently invite more automation and, thus, an elevated risk to ignore vital context in content moderation, in addition to the embedded limitations of automation.

While most content moderation decisions in our sample were enforced uniformly across the EEA and EU–accounting for 99.13% of all SoRs–our findings also point to a territorialised aspect of platform moderation, particularly in member states like the Netherlands, France, Italy and Spain. These countries consistently appeared in our data as regions where platforms applied territorialised content moderation decisions and that were more likely to include manual means of moderation. However, this territorial variation does not necessarily imply a legal fragmentation. Rather, it suggests that platforms are more selective about the localisation of their content moderation strategies within the overarching harmonised regulatory environment established by the DSA. Additionally, this selective territorialisation of content moderation–namely concerning certain aforementioned member-

states – shown by our findings indicate a tiered approach to content moderation that not only has to do with national legal frameworks and authorities but, potentially, also with how large a market is and, subsequently, how many resources are invested in the form of human reviewers speaking the local language.

Our analysis revealed a notable correlation between territorialised content moderation and the manual detection of flagged content, particularly through third-party notifications. Put simply, when content is moderated at the level of specific member-states rather than uniformly across the EU or EEA, it is more likely to have been flagged manually (e.g., through reports by users or other third-parties) rather than detected through automated systems. Under Article 16 of the DSA, platforms are required to implement user-friendly mechanisms enabling individuals to report potentially illegal content. However, in our dataset for TikTok, X, and YouTube, the majority of flagged content fell under the category of "other types of notifications." This vague classification does not provide details about the origin of these notifications, whether from individuals, organisations, or other entities.

Also, platforms are not obligated under DSA Article 9 to issue Statements of Reasons (SoRs) for content removed at the request of judicial or administrative authorities. This omission restricts our ability to differentiate between notifications stemming from public authorities versus other third parties in the context of the Transparency Database. However, a closer examination of the types of content flagged by third parties (Table 5) offers some insights. For instance, intellectual property violations overwhelmingly dominate YouTube's moderated content, underscoring the influence of rightsholders on platform governance. On the other hand, TikTok and X show broader diversity in flagged categories, including harmful speech, fraud and privacy violations. This suggests that different platforms cater to varying types of stakeholders' demands and observability can help us foreground these latent power dynamics.

At any rate, our study shows that we are moving towards a harmonised framework of platform governance in the EU. This harmonisation, however, is largely enforced by platforms as the "delegated enforcers" under the DSA (Husovec, 2024). As Kaushal et al. (2024) showed in their study of the Transparency Database, most of the content moderated is found to be incompatible with the platforms' terms of services and/or community guidelines rather than the basis of national or EU law. As a result, this delegation and interpretation of the DSA, raises broader questions about the power of platforms broadly and, particularly, the diversity and vibrancy of the EU's digital public spheres. For instance, how can we ensure that platforms do not disproportionately silence marginalised voices when they have such discre-

tion over the operationalisation of EU's regulations?

This delegation, moreover, of enforcement responsibilities to platforms, while practical from a regulatory perspective to ensure a more-or-less harmonised and "predictable" regulatory environment (Husovec & Roche Laguna, 2022), raises important concerns about the prioritisation of monocultural and compliant speech (Douek, 2020; Keller, 2024) and the potential cultivation of a sanitised digital public sphere (Griffin, 2022). Importantly, as Keller has repeatedly noted (2022, 2024), the DSA might end up favouring incumbent platform firms that have put in place an industrialised content moderation structure, while disproportionately affecting smaller platforms or discourage the experimentation with other systems of moderation, like more community-oriented or artisanal approaches (Caplan, 2018).

Automation, to be sure, plays a pivotal role in facilitating platforms' compliance with the DSA. In that sense, as demonstrated in our analysis of TikTok, industrial-scale automation is not only necessary for the industrial-scale of platforms' content circulation (Gillespie, 2020) but also streamlines processes of "automated transparency" (Kaushal et al., 2024) like the Transparency Database. This dynamic reflects a broader trend toward the *industrialisation of content moderation*, where platforms adopt algorithmic systems (Gorwa et al., 2020), create specialised Trust and Safety teams (Caplan, 2023) and implement transparency mechanisms to comply with regulations.

All these elements give way to a "factory-like" approach to content moderation (Keller, 2024), reconfiguring platforms into compliance-driven entities that prioritise operational efficiency over nuanced content moderation, precisely due to the systemic shortcomings and problems plaguing automation. We also find that the way that the DSA operationalises platform observability does not leave much space either for the kind of technical modifiability that Ferrari et al. (2023) consider to be crucial for digital governance or for experimentations with alternative content moderation systems. To be clear, these obligations formally apply only to designated VLOPs. However, the normative gravitas of the DSA risks reinforcing industrialised, centralised moderation approaches across the digital ecosystem. As such, the DSA risks further entrenching large platforms' power by restricting the potential of having a more plethoric and diverse content moderation ecosystem which would be more likely to be cultivated by smaller or decentralised platforms, which may now be pressured to align with industrial approaches to content governance (e.g., BlueSky's approach to a 'stackable' content moderation system; The BlueSky Team, 2024).

It may, moreover, depoliticise content governance, alongside the process of making it legible and accountable, allowing platforms to hide behind a discursive framing of content moderation as the product of neutral, automated systems that are the only option to comply with the DSA, rather than as decisions embedded in complex sociopolitical and cultural contexts (Ananny & Crawford, 2016). Therefore, while the industrialisation of content moderation has become unavoidable due to the scale and complexity of digital platforms, its implications for platform governance demand critical scrutiny, that is our capacity to observe and probe into these systems.

Implications for platform observability

In the context of this paper, where we focus on the DSA, observability translates to the systematic and –wherever possible– real-time access to platforms' data and the sociotechnical systems that allow for the production of that data and the overarching ecosystems. In other words, it refers to creation of the institutional and technical conditions necessary for the sustained, collaborative and situated observation of platforms. The Transparency Database introduced by the DSA partially responds to these calls for platform observability. Returning to Rieder and Hofmann's conceptualisation, we show that the DSA, here through the Transparency Database, contributes to the normative condition by mandating transparency practices that aim to serve public scrutiny. It also aligns with the sociotechnical condition by enabling public access to structured and real-time data, though with well-documented limitations. Last, the governance condition is also arguably met provided that the Database can be accessed and used by a wide variety of stakeholders who may collectively participate in governance processes.

However, once we consider Ferrari and colleagues' governance conditions (2023), we understand that the Transparency Database partially meets the condition of industrial observability as it provides some insights into the labour relations supporting platforms' content governance but there are significant inconsistencies and shortcomings. It also partially meets public inspectability as it makes a vast trove of data and insights available but primarily focuses on decisions of platforms' automated decision-making systems and not so much on more opaque processes (e.g., the models undergirding content governance). Last, technical modifiability is absent as there is no indication of how moderation systems can be reconfigured in response to observed shortcomings beyond regulatory proceedings and potential financial penalties. In that sense, these SoRs provide us with justifications for the platforms' decisions rather than explain how these decisions were made and car-

ried out or empower the involved parties to take further action (Leerssen, 2024, p. 25).

Regardless, we hold that this newfound level of access to important data might improve our chances for reigning in platform firms' unchecked power, primarily through ex-post accountability (e.g., fines). But given the unprecedented level of entanglement of platform infrastructures in our lives and political-economic systems this might no longer be – if it ever was – enough. For example, as mentioned earlier, observability should expand from just the consumer-facing (i.e., downstream) part of platforms' value chains to include more hidden layers (i.e., upstream) like the labour conditions of their outsourced contractors and environmental impact of their material infrastructures (Terzis, 2023).

Recent stories like the layoff of 300 human moderators from the Amsterdam office of TikTok (Nijssen, 2024) hint at a doubling down on the technical aspect of moderation, namely automation, potentially exacerbating the shortcomings and problems of automation discussed earlier in this paper. We must not forget that platform firms, and their content moderation systems, are also large employers operating on human labour. In that sense, conditions of observability should also extend to include the working spaces and the relations of production of platform capitalism (Srnicek, 2017). Doing so should not only allow us to observe and, ideally, ensure that these moderators work under decent conditions but also to better understand the value chains of content moderation and how platforms coordinate their workforce globally and, in so doing, create new opportunities for interventions. We understand that such a condition cannot be solely addressed by platform regulation but may also need more traditional regulatory frameworks moored in labour law (e.g., content moderators might need a similar instrument to the EU's platform work Directive).

Conclusion

This paper set out to explore how the DSA's Transparency Database operationalises the concept of platform observability. Building on a literature review of platform governance scholarship, we contextualised our paper with a conceptualisation of platform observability, primarily following Rieder and Hoffman (2023)'s framework and complementing it with Ferrari et al. (2023)'s governance conditions. Empirically, we analysed nearly 439 million SoRs from eight digital VLOPs, with a deeper dive into three (X, YouTube and TikTok), aiming to understand what kind of insights related to observability we could gain from this novel regulatory instrument. We

identified three key findings: first, moderation decisions are overwhelmingly applied uniformly across the EEA or EU; second, only a few platforms report engaging in territorially specific moderation and in markedly different ways; and third, reliance on automation correlates with faster enforcement timelines and uniform application of decisions, whereas manual moderation activities are associated with longer delays and more territorial-specific application.

Despite identified shortcomings, we join other scholars studying this emerging strand of scholarship (Kaushal et al., 2024; Drolsbach & Pröllochs, 2024; Trujillo et al., 2025) in recognising the current and potential benefits of the DSA and its Transparency Database for platform observability, as well as the caveats stemming primarily from the self-reported nature of most of the data available. As a result, while the database will prove to be an important resource for further research, the insights we can draw from it alone require strong analytical reflexes from our side. Crucial to this point is that the Transparency Database may enable a network of stakeholders to contribute to a shared governance model through the production of knowledge based on platforms' data (Helberger et al., 2018). As such, it offers a unique opportunity to engage critically with the intricacies of algorithmic content governance at an industrial scale, but its impact will depend on the extent to which it inspires a broader commitment to meaningful transparency, collaborative observability and structural reform in platform governance.

Future studies should also consider the specificities and affordances of each platform when using these novel governance mechanisms, which was not the case for our paper. For example, a future study should compare similar platforms like video-sharing platforms to investigate differences concerning the use of automation in moderation.

References

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Are, C. (2022). The Shadowban Cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, *22*(8), 2002–2019. https://doi.org/10.1080/14680777.2021.1928259

Birchall, C. (2021). *Radical secrecy: The ends of transparency in datafied America*. University of Minnesota Press.

Bloch-Wehba, H. (2020). Automation in moderation. *Cornell International Law Journal*, *53*(41). https://ssrn.com/abstract=3521619

Caplan, R. (2018). *Content or context moderation? Artisanal, community-reliant, and industrial approaches*. Data & Society. https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf

Caplan, R. (2023). Networked platform governance: The construction of the democratic platform. *International Journal of Communication*, *17*, 3451–3472. https://ijoc.org/index.php/ijoc/article/view/20035

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 2056305120936636. https://doi.org/10.1177/2056305120936636

Dergacheva, D., Kuznetsova, V., Scharlach, R., & Katzenbach, C. (2023). *One day in content moderation: Analyzing 24 h of social media platforms' content decisions through the DSA transparency database*. Universität Bremen. https://doi.org/10.26092/ELIB/2707

Dijck, J., Poell, T., & Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Douek, E. (2020). Governing online speech: From 'posts-as-trumps' to proportionality and probability. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3679607

Drolsbach, C. P., & Pröllochs, N. (2024). Content moderation on social media in the EU: Insights from the DSA transparency database. *Companion Proceedings of the ACM Web Conference 2024*, 939–942. https://doi.org/10.1145/3589335.3651482

European Union. (2022). *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) (Text with EEA Relevance)*. http://data.europa.eu/eli/reg/2022/2065/oj/eng

Ferrari, F., Van Dijck, J., & Van Den Bosch, A. (2025). Observe, inspect, modify: Three conditions for generative AI governance. *New Media & Society*, *27*(5), 2788–2806. https://doi.org/10.1177/14614448231214811

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

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

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234. https://doi.org/10.1177/2053951720943234

Gorwa, R. (2019a). The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1407

Gorwa, R. (2019b). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R. (2024). *The politics of platform regulation: How governments shape online content moderation* (First). Oxford University Press.

Gorwa, R., & Ash, T. G. (2019). *Democratic transparency in the platform society*.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1),

205395171989794. https://doi.org/10.1177/2053951719897945

Griffin, R. (2022). The sanitised platform. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4007098

Griffin, R. (2023). Public and private power in social media governance: Multistakeholderism, the rule of law and democratic accountability. *Transnational Legal Theory*, *14*(1), 46–89. https://doi.org/10.1080/20414005.2023.2203538

Grimmelmann, J. (2017). *The virtues of moderation*. LawArXiv. https://doi.org/10.31228/osf.io/qwxf5

Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, *8*(6), 842–854. https://doi.org/10.1080/21670811.2020.1773888

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Helberger, N., & Samuelson, P. (2024). *The Digital Services Act as a global transparency regime*. https://doi.org/10.59704/06c97b13f47ed11c

Husovec, M. (2024). Introduction: Taming the powers. In M. Husovec, *Principles of the Digital Services Act* (1st edn, pp. 3–18). Oxford University Press. https://doi.org/10.1093/law-ocl/9780192882455.003.0001

Husovec, M., & Roche Laguna, I. (2022). Digital Services Act: A short primer. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4153796

Jiménez-Durán, R. (2023). The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4590147

Kaushal, R., Van De Kerkhof, J., Goanta, C., Spanakis, G., & Iamnitchi, A. (2024). Automated transparency: A legal and empirical analysis of the Digital Services Act transparency database. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1121–1132. https://doi.org/10.1145/3630106.3658960

Keller, D. (2022). The DSA's industrial model for content moderation. *Verfassungsblog: On Matters Constitutional*. https://doi.org/10.17176/20220224-121133-0

Keller, D. (2024). The rise of the compliant speech platform. In *Lawfare*. https://www.lawfaremedia.org/article/the-rise-of-the-compliant-speech-platform

Keller, D., & Leerssen, P. (2020). Facts and where to find them: Empirical research on internet platforms and content moderation. In N. Persily & J. A. Tucker (Eds), *Social media and democracy* (1st edn, pp. 220–251). Cambridge University Press. https://doi.org/10.1017/9781108890960.011

Klonick, K. & P. (2018). The new governors: The people, rules, and processes governing online speech. *HARVARD LAW REVIEW*, *131*, 73 ,.

Leerssen, P. (2024). Outside the black box: From algorithmic transparency to platform observability in the Digital Services Act. *Weizenbaum Journal of the Digital Society*, *4*(2). https://doi.org/10.34669/WI.WJDS/4.2.3

Nicholas, G., & Bhatia, A. (2023). *Lost in translation: Large language models in non-english content analysis*. Center for Democracy & Technology. https://cdt.org/insights/lost-in-translation-large-lang

uage-models-in-non-english-content-analysis/

Nijssen, T. (2024, October 15). *Met ontslag Amsterdams moderatieteam wil TikTok meer werken met AI*. nrc. https://www.nrc.nl/nieuws/2024/10/15/met-ontslag-amsterdams-moderatieteam-wil-tiktok-meer-werken-met-ai-hoe-modereert-het-techbedrijf-a4869447

Papaevangelou, C. (2021). The existential stakes of platform governance: A critical literature review. *Open Research Europe*, *1*, 31. https://doi.org/10.12688/openreseurope.13358.2

Poell, T., Nieborg, D. B., & Duffy, B. E. (2021). *Platforms and cultural production*. Polity Press.

Posada, J. (2022). Embedded reproduction in platform data work. *Information, Communication & Society*, *25*(6), 816–834. https://doi.org/10.1080/1369118X.2022.2049849

R.Core Team. (2021). *R: A language and environment for statistical computing, Vienna [Computer software]*. R Foundation for Statistical Computing. https://www.R-project.org

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

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Srnicek, N. (2017). *Platform capitalism*. Polity.

Terzis, P. (2023). Law and the political economy of AI production. *International Journal of Law and Information Technology*, *31*(4), 302–330. https://doi.org/10.1093/ijlit/eaae001

Terzis, P., Veale, M., & Gaumann, N. (2024). Law and the emerging political economy of algorithmic audits. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1255–1267. https://doi.org/10.1145/3630106.3658970

Trujillo, A., Fagni, T., & Cresci, S. (2025). The DSA transparency database: Auditing self-reported moderation actions by social media. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1–28. https://doi.org/10.1145/3711085

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society*, *10*(1), 20539517231172424. https://doi.org/10.1177/20539517231172424

Zornetta, A. (2024, July 11). *Is The Digital Services Act truly a transparency machine?* Tech Policy Press. https://techpolicy.press/is-the-digital-services-act-truly-a-transparency-machine

Appendix

TABLE 4: Top ten content moderation decisions with territorial scope by YouTube, Tiktok and X

PLATFORM	TERRITORIAL SCOPE	RANK	SORS	% SORS	SORS/AMAR
TIKTOK	EEA (NO ICELAND)	1	221M	50.40%	1.03524822
TIKTOK	EEA	2	213M	48.41%	1.02165005
TIKTOK	EU	3	1M	0.32%	1.21562522
TIKTOK	DE	4	588K	0.13%	0.802892
TIKTOK	FR	5	432K	0.10%	0.75378213
TIKTOK	IT	6	323K	0.07%	0.51555205
TIKTOK	PT	7	197K	0.04%	1.26603687
TIKTOK	PL	8	185K	0.04%	0.97310199
TIKTOK	ES	9	169K	0.04%	1.02537825
TIKTOK	FI	10	129K	0.03%	1.58220212

PLATFORM	TERRITORIAL SCOPE	RANK	SORS	% SORS	SORS/AMAR
YOUTUBE	NL	1	43K	0.01%	0.7028192
YOUTUBE	FR	2	218K	0.05%	0.6018377
YOUTUBE	ES	3	234K	0.05%	0.6321022
YOUTUBE	PL	4	198K	0.04%	0.75431209
YOUTUBE	PT	5	19K	0.01%	0.62376882
YOUTUBE	IE	6	22K	0.01%	0.99028031
YOUTUBE	IT	7	232K	0.05%	0.89425723
YOUTUBE	FI	8	19K	0.01%	0.93103291
YOUTUBE	EU	9	29K	0.01%	0.50291721
YOUTUBE	AT	10	15K	0.00%	1.22055035

PLATFORM	TERRITORIAL SCOPE	RANK	SORS	% SORS	SORS/AMAR
X	NL	1	42K	0.01%	0.75768104
X	FR	2	218K	0.05%	0.6757983

PLATFORM	TERRITORIAL SCOPE	RANK	SORS	% SORS	SORS/AMAR
X	ES	3	234K	0.05%	0.73201781
X	PL	4	198K	0.04%	0.90390533
X	PT	5	19K	0.01%	0.74893766
X	IE	6	22K	0.01%	1.0306261
X	IT	7	232K	0.05%	0.92168241
X	FI	8	19K	0.01%	1.15494221
X	EU	9	29K	0.01%	0.60002643
X	AT	10	15K	0.00%	1.25769104

TABLE 5: Categories of content detected by third-parties

PLATFORM NAME	ANIMAL WELFARE	RISK FOR PUBLIC SECURITY	NON-CONSENSUAL BEHAVIOUR	NEGATIVE EFFECTS ON CIVIC DISCOURSE OR ELECTIONS	SELF-HARM	INTELLECTUAL PROPERTY INFRINGEMENTS	PROTECTION OF MINORS	DAT PROTEC ANI PRIVA VIOLAT
TikTok	0	748	2.494	4.329	1.501	5.352	265.062	4.652
X	0	27	386	3.539	2	7.391	2.311	34.973
YouTube	0	0	0	0	0	3.515.451	28	6.404

TABLE 6: Human reviewers speaking EU's official languages employed by YouTube, X and TikTok

EU OFFICIAL LANGUAGE	YOUTUBE	X	TIKTOK
Bulgarian	9	2	69
Croatian	24	1	20
Czech	31	0	62
Danish	9	0	42
Dutch	24	1	167
English	142	2,294	2137
Estonian	7	0	6
Finnish	15	0	40
French	176	52	687

EU OFFICIAL LANGUAGE	YOUTUBE	X	TIKTOK
German	231	81	869
Greek	28	0	96
Hungarian	25	0	63
Irish	0	0	439
Italian	91	2	0
Latvian	11	1	9
Lithuanian	11	0	6
Maltese	0	0	0
Polish	99	1	208
Portuguese	464	41	75
Romanian	34	0	167
Slovak	5	0	44
Slovenian	15	0	45
Spanish	507	20	468
Swedish	16	0	108
Total	**1974**	**2496**	**5827**

Published by
Alexander von Humboldt Institute for Internet and Society
CAIS Research for the Digital Age
in cooperation with
CREATe
cnrs centre internet et societe
R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya
University of Tartu Johan Skytte Institute of Political Studies