VOLUME 6 March 2025 ISSUE 1

JOURNAL ON EMERGING TECHNOLOGIES



ARTICLES

# CLEARING THE HAZE: IS THE EU DIGITAL SERVICES ACT FINALLY FORCING PLATFORMS TO OPEN UP ABOUT CONTENT MODERATION?

*ALESSIA ZORNETTA*

Abstract

This article presents a comparative analysis of transparency reporting practices by Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) in the context of the Digital Services Act (DSA), aiming to assess the legislation's impact on enhancing transparency in online platforms' operations. The DSA, a landmark EU regulation, mandates rigorous transparency obligations to promote accountability. Through an examination of transparency reports from 19 VLOPs and VLOSEs, this article reveals the changes induced by the DSA transparency mandates. Despite the regulatory push for enhanced transparency, the findings suggest that post-DSA transparency reports continue to fall short in providing the depth, breadth, and quality necessary for effective cross-comparison and scrutiny. The article proposes specific implementation measures to address the gaps left by the DSA as well as recommendations for DSA-like regulatory efforts outside of the European Union.

# Clearing the Haze: Is the EU Digital Services Act Finally Forcing Platforms to Open Up About Content Modernization?

*ALESSIA ZORNETTA*

## INTRODUCTION

"Online platforms are at the core of some of the key aspects of our daily lives, democracies, and economies. It's only logic that we ensure that these platforms live up to their responsibilities in terms of reducing the amount of illegal content online and mitigating other online harms, as well as protecting the fundamental rights and safety of users," remarked Margrethe Vestager, the EU Executive Vice-President for a Europe Fit for the Digital Age on November 16th, 2022, when the Digital Services Act came into force.[1] The Digital Services Act aims to be a landmark regulation attempting to address the societal risks posed by online platforms through transparency, accountability, and user empowerment.[2]

The DSA marks a significant milestone in the European Union's efforts to regulate online platforms. Among its many provisions, the DSA introduces transparency obligations for online platforms to promote greater accountability and oversight of their content moderation practices. Meaningful transparency is essential for promoting accountability and oversight of online platforms. Transparency allows users to better understand how platforms moderate content, decide what content to promote or demote, and how algorithms shape their online experience. Transparency can also help regulators, researchers, and civil society organizations scrutinize platforms' practices, identify potential biases and errors, and hold platforms accountable for their decisions.

This article undertakes a comprehensive analysis of transparency reporting practices of 19 Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) before and after the transparency obligations of the DSA entered into force. Before delving into the comparative

---

[1] European Commission Press Release, Digital Services Act: EU's Landmark Rules for Online Platforms Enter into Force (Nov. 16, 2022), https://ec.europa.eu/commission/presscorner/detail/en/ip_22_6906.

[2] *See* 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 O.J. (L 277), http://data.europa.eu/eli/reg/2022/2065/oj/eng.

analysis, the article highlights the critical role of transparency and its different shapes and provides an overview of the obligations found in the DSA and its enforcement framework. The article concludes with an assessment of a key feature of transparency mandates: standardization.

## I. MEANINGFUL TRANSPARENCY IN CONTENT MODERATION PRACTICES

Online platforms engage in the practice of "content moderation" to regulate user activity. While each platform has its own unique definition of "content moderation," the broad consensus is that through this practice, user-generated content is either promoted or demoted, uploaded or removed, and monetized or demonetized. Platforms are governed by specific rules and legal obligations that users must comply with, such as community guidelines and terms of service. In addition, platforms employ internal decision-making processes to develop automated content moderation systems, escalation procedures, and algorithmic ranking. The outcome of content moderation decisions can profoundly impact the visibility and accessibility of content, which, in turn, can shape public discourse.[3]

However, content moderation is a field shrouded in opacity, with platforms retaining the discretion to determine what information to disclose to the public and how. Currently, the only available information regarding content moderation arises from platforms' own disclosures, investigative efforts undertaken by journalists and researchers, and leaked documents by whistleblowers.

### A. Transparency as a Vector for Compliance

The idea of using transparency as a tool to promote accountability and trust is not a novelty of platform governance studies. Transparency and the "right to know" are founding principles of democratic systems.[4]

Modern conceptions of transparency have been linked to the empowerment of citizens through access to information, thus facilitating government accountability. [5] In the private sector, transparency

---

[3] *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018); Evelyn Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMEND. INSTIT. COLUM. UNIV. (2020), https://papers.ssrn.com/abstract=3572309.

[4] Archon Fung, *Infotopia: Unleashing the Democratic Power of Transparency*, 41 POL. & SOC'Y 183 (2013).

[5] *See* Mikkel Flyverbom, *Sunlight in Cyberspace? On Transparency as a Form of Ordering*, 18 EUR. J. SOC. THEORY 168 (2015),

initiatives have influenced internal operations by shaping public opinion and improving regulatory oversight.[6] The primary assumption is that, by making available to the public scrutiny the internal decision-making processes of corporations or governments, customers and citizens will be able to make informed decisions. Additionally, for corporate transparency, the assumption is further complemented by the idea that, by accessing corporate information, regulators will be able to ensure compliance with legal requirements and intervene where self-regulation fails.[7]

In the digital age, transparency promotes accountability and trust in platforms content moderation processes--it is a necessary component within a system of accountability."[8]

The increasing impact of content moderation decisions in shaping public discourse and the opacity characterizing such processes has forced calls for greater transparency by the general public, activists, academics, and a multitude of regulators worldwide.[9] The role platforms play in organizing content's visibility, and thus defining the paths of online (and offline) speech and access to information, justifies such calls. Nevertheless, calls for greater transparency in content moderation have suffered from a lack of specificity, allowing platforms to exploit transparency mechanisms to avoid oversight.[10]

To be effective, transparency must be meaningful. Meaningful transparency is targeted to specific goals and thus structured to address specific needs.[11] To achieve the goal of accountability, the different receivers of the information disclosed need to be able to interpret and

---

https://journals.sagepub.com/doi/10.1177/1368431014555258 (last visited Jan. 6, 2023); Oana Brindusa Albu & Mikkel Flyverbom, *Organizational Transparency: Conceptualizations, Conditions, and Consequences*, 58 BUS. & SOC'Y 268 (2019).

[6] *See generally* Don Tapscott & David Ticoll, *The Naked Corporation: How the Age of Transparency Will Revolutionize Business*, FREE PRESS(2003).

[7] *See* Alessia Zornetta, *Online Misinformation: Improving Transparency in Content Moderation Practices of Social Media Companies*, MCGILL UNIV. (2022), https://escholarship.mcgill.ca/concern/theses/rf55zd782; Daphne Keller & Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation, in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM 220, 224 (Nathaniel Persily & Joshua A. Tucker eds., 2020); Fung, *supra* note 4.

[8] Nicolas P. Suzor et al., *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMMC'N 1526, 1527 (2019).

[9] Sonja Solomun, Maryna Polataiko & Helen A. Hayes, *Platform Responsibility and Regulation in Canada: Considerations on Transparency, Legislative Clarity, and Design*, 34 HARV. J.L. & TECH. (2021).

[10] Suzor et al., *supra* note 8.

[11] *Cf. Id.*

understand what the information is communicating.[12] When meaningful disclosure empowers receivers of such information to hold the disclosers accountable by modifying their behavior, it forces a change in decision-making.[13]

### B. Transparency's Addressees

At present, voluntary transparency practices by online platforms have shown a variety of weaknesses in achieving meaningfulness.[14] The tendency to focus on aggregate statistical data prevents addressees of transparency to observe the entirety of the content moderation process. Additionally, the inability to access information around the context in which content moderation decisions are made prevents the understanding of what types of content are primarily impacted and overall patterns and trends.[15] Current voluntary transparency practices leave the processes, protocols, and procedures that lead to internal policy and decision-making outside their scope.[16]

When it comes to online platforms, transparency can take different shapes and forms depending on to whom it is addressed and the goal of disclosing information. The following subsections provide an overview of different transparency shapes, what information would be relevant (in an ideal scenario), and how they benefit their receivers.

#### 1. Transparency Towards Directly Affected Users

When users interact within online platforms, specific rules guide the user behavior and define what content is or is not allowed. Platforms set the general rules for user activity in their Terms of Service (ToS). ToS establish a legal agreement between a platform and the user, not only regulating user behavior but also defining rights and responsibilities,

---

[12] Albu & Flyverbom, *supra* note 5. *See generally* Archon Fung, Mary Graham & David Weil, FULL DISCLOSURE: THE PERILS AND PROMISE OF TRANSPARENCY (2007).
[13] Zornetta, *supra* note 7.
[14] *Cf.* Eleni Kosta & Magdalena Brewczyńska, *Government Access to User Data: Towards More Meaningful Transparency Reports*, *in* REGULATING INDUSTRIAL INTERNET THROUGH IPR, DATA PROTECTION AND COMPETITION LAW (Ballardini, Kuoppamäki & Pitkänen eds. 2019).
[15] Solomun, Polataiko & Hayes, *supra* note 9; Chris Tenove & Heidi Tworek, *Processes, People and Public Accountability* (2020), https://ppforum.ca/articles/processes-people-and-public-accountability/ (last visited Apr 26, 2023).
[16] Robert Gorwa & Timothy Garton Ash, *Democratic Transparency in the Platform Society*, SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM 286 (2020).

privacy policies, and intellectual property rights. Most platforms supplement their ToS with Community Standards,"—also named Community Guidelines,"[17] "Acceptable Use Policy,"[18] or Rules."[19]

Platforms engage in content moderation to enforce ToS, Community Standards, and local laws affecting content and user behavior online. Content moderation can impact user activity in a variety of ways. When it affects an individual piece of content, it could lead to content removal, demotion, or, if applicable, demonetization. When it affects the overall user behavior, it could lead to temporary account suspension, permanent removal, or even the blocking of newly created accounts belonging to recurring violators. Different platforms develop different processes to decide what thresholds must be met to trigger compliance and what resulting measures will be applied.

In this context, platform transparency refers to the provision of an explanation to users whose content or account has been involved in a content moderation decision leading to one of the actions mentioned above. Platforms provide different degrees of specificity in user explanations. In an ideal scenario, users would be made aware of the policy or law infringed on by their content or activity, the facts taken into account when establishing the infringement, whether the decision was made by a human moderator or via automated tools, and the available options for redressal.[20] The notification received by the user should also contain specific details of the affected content.[21]

The goals of providing such information can be twofold: educating users and legitimizing content moderation decisions. First, affected users who receive an explanation can verify whether the decision was accurate and eventually learn from such experience to avoid repeating the same infringement. Alternatively, in case of a wrongful decision, users can (ideally) submit an appeal, provide additional contextual information, and potentially reinstate their content or account. Second, platforms can establish trust with their users and demonstrate that content moderation is performed fairly and consistently and that there is a specific process for deciding on infringing content or activity.

---

[17] *See, e.g.*, *Community Guidelines*, TIKTOK, https://www.tiktok.com/community-guidelines/en/ (last visited Apr. 24, 2023).
[18] *See, e.g.*, *Shopify Acceptable Use Policy*, SHOPIFY, https://www.shopify.com/legal/aup (last visited Apr. 24, 2023).
[19] *See, e.g.*, The X Rules, X HELP CTR., https://help.twitter.com/en/rules-and-policies/twitter-rules (last visited Apr. 24, 2023).
[20] Suzor et al., *supra* note 8.
[21] *Id.* (arguing that individual notices should contain "URL of the prohibited content or a sufficiently detailed extract").

In an ideal scenario, users should be able to understand and interpret the notices received over moderation decisions affecting their accounts or content. Individual users should receive a notice not only when moderation leads to removal, but also when decisions result in demotion or reduced visibility.

However, the reality is that most users receive very little detail about content moderation decisions. At most, users are informed of the specific policy that their content or activity violated. Usually, no information is available regarding the facts and circumstances considered, or whether the decision was taken by a human moderator or via an automated tool. Although some platforms direct users to appeal portals, users usually complain that such mechanisms are often useless, unfair, and lead to no redress.[22]

In some cases, explanations are not provided at all, especially when the enforcement action is demotion.[23] Platforms use demotion in moderation processes when the content appears vaguely infringing but not to the point that it breaches a specific policy. Although demotion has comparable effects to removal, users are rarely notified and thus deprived of the possibility of redressal.[24]

Another aspect of platform transparency towards affected users regards transparency over users' flagging of another users' activity. User flaggers often do not receive any notification of the decision made after referring a piece of content or an account due to a possible infringement. In particular circumstances, such as when accounts impersonate others' identities, users who flag such infringement are left with no answer and no opportunity for appeal in case of denial.

2. Transparency Towards the Public

In its most popular connotation, platform transparency refers to transparency initiatives addressed to the general public. The practice of

---

[22] *Id.*; Kristen Vaccaro, Christian Sandvig & Karrie Karahalios, *"At the End of the Day Facebook Does What It Wants": How Users Experience Contesting Algorithmic Content Moderation*, 4 PROC. ACM HUM.-COMPUT. INTERACT. 167:1 (2020).
[23] *See* Paddy Leerssen, *An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation*, 48 COMPUT. L. & SEC. REV. 1 (2023); Callie Middlebrook, *The Grey Area: Instagram, Shadowbanning, and the Erasure of Marginalized Communities* (2020), https://papers.ssrn.com/abstract=3539721 (last visited Apr 24, 2023); Kelley Cotter, *"Shadowbanning Is Not a Thing": Black Box Gaslighting and the Power to Independently Know and Credibly Critique Algorithms*, 26 INFO., COMMC'N & SOC'Y 1226 (2021); Laura Savolainen, *The Shadow Banning Controversy: Perceived Governance and Algorithmic Folklore*, 44 MEDIA, CULTURE & SOC'Y 1091 (2022).
[24] Savolainen, *supra* note 23; Leerssen, *supra* note 23.

disclosing information on platforms' internal operations started in the early 2010s mainly as a commitment to inform the public on content takedowns and account information requests received from governmental agencies worldwide.[25] Especially during the aftermath of Edward Snowden's revelations, online platforms turned towards transparency reports to boost public trust and distance themselves from governments.[26]

Over time, transparency reports also evolved to shed light on companies' content moderation processes. In particular, the Cambridge Analytica scandals in 2016 fueled public demands for broader transparency from platforms. 2018 marked a turning point in voluntary transparency as major platforms began to release public-facing community standards and publish transparency reports on their enforcement.[27]

Transparency reports addressed to the general public vary significantly among online platforms in frequency, format, and content. Reports generally contain aggregate data about community guidelines enforcement across a specific platform. Some platforms also break down such aggregate data by location, while others distinguish between compliance with legal obligations and enforcement of platforms' own policies.

In addition to reports, some platforms also publish somewhat detailed explanations for policies and processes for enforcement. For example, Facebook's policy on "Dangerous Individuals and Organizations" contains a "policy rationale" where the company explains the reasoning behind their tri-partite designation process and their respective consequences. [28] Similarly, YouTube's "Misinformation policies" page provides users with examples of content not allowed on the platform and clarification on how the company handles violations following a three-strikes approach.[29]

---

[25] *See, e.g.*, David Drummond, *Greater Transparency around Government Requests*, GOOGLE BLOG (Apr. 20, 2010), https://googleblog.blogspot.com/2010/04/greater-transparency-around-government.html.
[26] Robert Gorwa & Timothy Garton Ash, *Democratic Transparency in the Platform Society*, SOC. MEDIA DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM 286, 295-297 (2020).
[27] *Id.*
[28] *Dangerous Organizations and Individuals*, META TRANSPARENCY CTR., https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (last visited Apr. 24, 2023).
[29] *Misinformation Policies*, YOUTUBE, https://support.google.com/youtube/answer/10834785?hl=en-GB&ref_topic=10833358 (last visited Apr. 24, 2023); *Community Guidelines Strike Basics on YouTube*, YOUTUBE, https://support.google.com/youtube/answer/2802032?sjid=14849294838613030406-NA (last visited Apr. 24, 2023).

Additionally, larger platforms tend to publish press statements with further clarifications on high-profile decisions. When major online platforms decided to deplatform former U.S. President Donald Trump on January 6th, 2021, such decisions came accompanied by detailed explanations in platforms' blogs.[30] Similarly, in the aftermath of the Christchurch Terrorist Attacks, Facebook issued a press statement explaining how the company was introducing restrictions to the "live" feature.[31]

Nevertheless, such transparency mechanisms have been long criticized for being primarily a PR-management tool rather than a resource aimed at reducing opacity and ensuring accountability.[32] Regarding transparency reports, the main criticism focuses on the lack of independent verification of platforms' data. By only disclosing data in aggregate form and significantly limiting independent review, it is impossible to confirm the disclosed information or obtain a comprehensive overview of platforms' approaches with different content and user activity.[33] Additionally, the use of different metrics by platforms renders empirical analysis of content moderation approaches particularly challenging.[34]

Another fundamental flaw of transparency reports is the lack of specificity.[35] Platforms tend to rely on broad and generic terminology and disperse information through various reports, thus undermining the identification of trends and procedural gaps.[36] Lastly, most (if not all) transparency reports highlight content takedowns and account suspensions while leaving any information on engagement, demotion, and other "soft" enforcement measures outside of their scope.

Transparency over platforms' policies and the decision-making process that led to adopting specific wording is a recent novelty that still has to demonstrate efficacy. Once more, this move is seen as a way to avoid regulatory oversight by giving the public the illusion of legitimacy.

Finally, transparency for high-profile cases does little to inform

---

[30] *See, e.g.*, Nick Clegg, *In Response to Oversight Board, Trump Suspended for Two Years; Will Only Be Reinstated if Conditions Permit*, META (June 4, 2021), https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/.

[31] Guy Rosen, *Protecting Facebook Live from Abuse and Investing in Manipulated Media Research*, META (May 14, 2019), https://about.fb.com/news/2019/05/protecting-live-from-abuse/.

[32] *See* Suzor et al., *supra* note 8.

[33] *See* Zornetta, *supra* note 7.

[34] Amélie Heldt, *Reading between the Lines and the Numbers: an Analysis of the First NetzDG Reports*, 8 INTERNET POL'Y REV. 1 (2019).

[35] Keller and Leerssen, *supra* note 7 at 221.

[36] *See* Zornetta, *supra* note 7.

the general public on how content moderation processes work. Although such explanations provide more detailed information about individual cases, leaked documents have revealed that platforms tend to distinguish their approach towards different categories of users. For instance, in the "Facebook Files," whistleblower Frances Haugen revealed that the company used a "CrossCheck" system to filter posts by known high-profile users and handle them with ad hoc procedures.[37] Therefore, high-profile press statements do little to shed light on platforms' approach to the majority of its content.

3. Transparency Towards Researchers

Another aspect of transparency lies in allowing independent researchers – broadly defined – to company's data, research, employees, and decision-making processes. Independent researchers have the ability to corroborate platforms' statements, without incurring conflicts of interest and limitations such as fiduciary duties towards shareholders.[38] The primary goal of transparency towards researchers is to verify companies' statements, identify industry trends through cross-comparison of different platforms, and, overall, studying the impact of digital platforms on society and human behavior. Additionally, researchers can support policymaking by drawing attention to different issues among platforms.[39]

In an ideal scenario, the term "researchers" should be understood in its broadest meaning, thus not only comprising those affiliated with academic institutions but also including independent researchers, journalists, and civil society actors.[40] Although some vetting is necessary to safeguard users' privacy, too high of vetting standards might preclude

---

[37] Dan Milmo, *Facebook: Some High-Profile Users 'Allowed to Break Platform's Rules,'* THE GUARDIAN (Sept. 13, 2021, 3:50 PM EDT), https://www.theguardian.com/technology/2021/sep/13/facebook-some-high-profile-users-allowed-to-break-platforms-rules.

[38] Nathaniel Persily, *A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act*, 1 J. ONLINE TR. & SAFETY (2021) at 1.

[39] *Id.*

[40] Laura Edelson, Inge Graef & Filippo Lancieri, *Access to Data and Algorithms: For an Effective DMA and DSA Implementation,* CTR. ON REGUL. IN EUR. (Mar. 2023); Richard Kuchta, Beatriz Almeida Saab & Lena-Maria Böswald, *The Data Access Problem: Limitations on Access to Public Data on Very Large Online Platforms*, DEMOCRACY REPORTING INT'L (Mar. 3, 2023), https://democracy-reporting.org/en/office/global/publications/the-data-access-problem-limitations-on-access-to-public-data-on-very-large-online-platforms; Caitlin Vogus, *Independent Researcher Access to Social Media Data: Comparing Legislative Proposals*, CTR. FOR DEMOCRACY & TECH. (Apr. 21, 2022), https://cdt.org/insights/independent-researcher-access-to-social-media-data-comparing-legislative-proposals/.

researchers' access, especially those from less privileged backgrounds. To effectively study and verify companies' claims, researchers should be given access to as much data as possible. The specific goals of each research project should be considered when selecting data, but, at a minimum, access should be allowed to all data that the platforms make available for sale on their commercial interfaces.[41]

However, platforms have routinely been reluctant to enable researchers' access. After the initial research momentum gained through the creation of Application Programming Interfaces (APIs) – where platforms would willingly allow researchers entrance into their "walled gardens" – platforms have recently moved away from APIs and began imposing more and more obstacles to researcher access.[42] Upon closing most APIs, platforms introduced specific vetting procedures to allow access to a handful of researchers at prestigious institutions. [43] Nevertheless, most of such initiatives were also abruptly ended or significantly restricted, both in terms of research tools and dissemination.[44]

4. Transparency Towards Regulators

Lastly, platform transparency can also be found in the shape of compliance reports resulting from mandatory obligations imposed by local laws. The primary goal of this aspect of transparency is that of informing legislators and regulators, as it promotes decisions based on accurate and complete information.[45] For example, being able to access data about the prevalence of specific categories of content, the total amount of content removed, and the underlying reasons for content moderation decisions might support policymakers in identifying areas for improvement and developing tailored policies. By providing policymakers with access to such information, transparency can bring light on the impact of such processes on society as a whole and point to

---

[41] *See* Persily, *supra* note 38.
[42] Alexander Halavais, *Overcoming Terms of Service: A Proposal for Ethical Distributed Research*, 22 INFO., COMMC'N & SOC'Y 1567 (2019).
[43] *See, e.g.*, Da Li et al., *Introducing the Researcher Platform: Empowering Independent Research Analyzing Large-Scale Data from Meta*, META RESEARCH (Jan. 11, 2022), https://research.facebook.com/blog/2022/1/introducing-the-researcher-platform-empowering-independent-research-analyzing-large-scale-data-from-meta/ (last visited May 4, 2023).
[44] For instance, Facebook's partnership with the Social Science Research Council only allowed researchers to use encrypted and locked computers within secure rooms at their headquarters. *See id.* at 1569*; See also* Kuchta, Saab, and Böswald, *supra* note 40.
[45] Persily, *supra* note 38.

areas where further regulatory action is needed.[46]

A well-known example of this kind of transparency are the compliance reports produced by covered platforms under the German Network Enforcement Act (NetzDG).[47] Under the NetzDG, companies with over two million users in Germany have to publish a bi-annual transparency report containing "general observations outlining the efforts undertaken by the social network provider to eliminate criminally punishable activity on the platform."[48]

Ideally, transparency mechanisms addressed to regulators should serve to improve regulatory scrutiny over platforms' impact on democratic processes.[49] Platforms should make available comprehensive information about content moderation processes, trends, and systems to mitigate harm. Where applicable, transparency should also serve to demonstrate compliance with local laws.[50] Regulators should be able to verify companies' claims independently.[51]

Nevertheless, previous examples have demonstrated the weaknesses of compliance reports. Without access to granular information about trends, behavior, and recurring incidents, policymakers are left to take action unsupported by empirical evidence. While aggregate statistics can offer a general understanding of content moderation practices and highlight broad areas of concern, they do not provide enough detail to hold platforms accountable through a comprehensive analysis.[52]

---

[46] Solomun, Polataiko, & Hayes, *supra* note 9; Nicolas P. Suzor et al., *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. OF COMMC'N 1526 (2019); *See also DSA*, Recital 66.

[47] *See* similar attempts of demanding transparency in France and Singapore: Loi organique n° 2018-1201 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information. JO, 22 Dec. 2018 (French law against manipulation of information); Protection from Online Falsehoods and Manipulation Bill, (Government Gazette, Notification No. B 10, 1 April 2019).

[48] Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG) German Law Archive, https://germanlawarchive.iuscomp.org/?p=1245 (last visited Jan. 6, 2023).

[49] Mike Ananny & Kate Crawford, *Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973 (2018), https://journals.sagepub.com/doi/epub/10.1177/1461444816676645 (last visited Apr. 11, 2023).

[50] *Cf.* Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG) German Law Archive, *supra* note 48; Heldt, *supra* note 34; Ben Wagner et al., *Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act*, in PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 261 (2020), https://dl.acm.org/doi/10.1145/3351095.3372856 (last visited Apr. 11, 2023).

[51] *DSA*, Recital 137.

[52] Suzor et al., *supra* note 8.

Exemplary of such issue are the findings of two studies which found "no evidence of a meaningful relationship between exposure to the Russian foreign influence campaign and changes in attitudes, polarization, or voting behavior."[53] For many years, lawmakers have called for increased regulation of political advertising online, even leading to specific laws being passed in Canada and France – calling for a temporary ban on online advertisement during specific election periods. Recognizing the need to contextualize the results of the two studies, commentators have pointed to how they shed light on how overestimated the impact of social media advertisement on voters might have been.[54] Without access to platforms' data, policymakers risk concentrating legislative and regulatory efforts on the wrong issue.[55]

## II. THE DIGITAL SERVICES ACT

The Digital Services Act is a regulation aimed at changing the dynamics of online service providers in the EU in an unprecedented manner. It belongs to a series of regulations and directives included in the project titled "A Europe fit for the digital age: Empowering people with a new generation of technologies."[56] The package also includes the Digital Markets Act,[57] an EU Cybersecurity Strategy, the proposed AI Act,[58] many other legislative initiatives, and partnerships with third countries[59].

The goal of the DSA is to "contribute to the proper functioning of

---

[53] Gregory Eady et al., *Exposure to the Russian Internet Research Agency Foreign Influence Campaign on Twitter in the 2016 US Election and Its Relationship to Attitudes and Voting Behavior*, 14 NATURE COMMC'NS 62 (2023).
[54] Evelyn Douek & Alex Stamos, *MC Weekly Update 1/16: Looking at the Evidence*, MODERATED CONTENT (Jan. 17, 2023), https://law.stanford.edu/podcasts/mc-weekly-update-1-16-looking-at-the-evidence/.
[55] *Cf.* Suzor et al., *supra* note 8.
[56] *A Europe Fit for the Digital Age*, EUR. COMM'N (2020), https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age_en (last visited Apr. 22, 2023).
[57] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance), 2022 O.J. (L 265) 1, http://data.europa.eu/eli/reg/2022/1925/oj/eng (last visited Dec. 16, 2022).
[58] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (last visited Apr. 22, 2023).
[59] *See, e.g.*, European Commission Press Release, EU and Singapore Launch Digital Partnership, (Feb. 1, 2023), https://ec.europa.eu/commission/presscorner/detail/en/ip_23_467 (last visited Apr. 22, 2023).

the internal market for intermediary services by setting out harmonised rules for a safe, predictable and trusted online environment that facilitates innovation and in which fundamental rights enshrined in the Charter [...] are effectively protected." [60] The DSA does not aim at controlling new technologies or mandating what content can or cannot be uploaded and hosted by online platforms. Instead, it focuses on minimizing risk of harm generated by and through online platforms. It does so by improving transparency, increasing oversight, demanding that platforms identify and mitigate risks their services might pose on society, and empower users.

To avoid curbing competition by imposing too many restrictions – technically and financially – the DSA imposes asymmetrical duties and obligations on different kinds of intermediaries. [61] The regulation distinguishes among intermediary services, hosting services, online platforms, and very large online platforms (VLOPs) and search engines (VLOSEs).

The decision to develop an asymmetrical regulation lies in the need to adapt obligations to the "type, size and nature of the intermediary service concerned" to facilitate the functioning of the internal market and ensure that different public policy objectives are achieved.[62] This article focuses primarily on the obligations imposed on VLOPs and VLOSEs, as they must comply with all the different obligations set out in the DSA. The European Commission is tasked with the designation of "VLOP" or "VLOSE."[63] According to article 33, platforms or search engines with a number of average monthly users of at least 45 million in the EU must comply with the strictest obligations. Online service providers must publish a report on their active monthly users every six months.[64]

### A. Transparency Reporting Obligations

---

[60] *DSA*, art. 1(1) & Recital 9.

[61] Bruna Martins dos Santos & David Morar, *Four Lessons for U.S. Legislators from the EU Digital Services Act*, BROOKINGS (Jan. 6, 2021), https://www.brookings.edu/blog/techtank/2021/01/06/four-lessons-for-u-s-legislators-from-the-eu-digital-services-act/; Solomun, Polataiko & Hayes, *supra* note 9.

[62] *DSA*, Recitals 40 & 41.

[63] *DSA*, art. 33.

[64] *Id.* art. 24(2). At the time of writing, the EU Commission has designated the following services as VLOPs: Ali Express, Amazon Store, Apple AppStore, PornHub, Booking.com, Facebook, Google Play, Google Maps, Google Shopping, Instagram, LinkedIn, Pinterest, Snapchat, Stripchat, TikTok, Wikipedia, X, XVideos, YouTube, and Zalando. The following services were designated as VLOSEs: Bing and Google Search. *See* European Commission, *Supervision of the Designated Very Large Online Platforms and Search Engines under DSA* (2024), https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses (last visited Feb. 26, 2024).

The DSA imposes a variety of transparency obligations on online service providers, including specific requirements for Terms & Conditions (Article 14), Individual Statements of Reasons (Article 17), and Data Access (Article 40). Considering the limited scope of this article, this section will focus on the requirements on transparency reporting requirements under Article 15 and Article 42.

The transparency reporting obligations follow the asymmetrical outline of the Digital Services Act, demanding greater disclosure from VLOPs. Article 15 applies to all online service providers, regardless of size, revenue or userbase, except for small or micro enterprises.[65] Article 24 supplements article 15 with additional reporting obligations for providers of online platforms, and Article 42 adds obligations for VLOPs and VLOSEs. Lastly, "trusted flaggers" are also obligated to publish a transparency report, although the discussion of the role played by trusted flaggers is beyond the scope of this article.[66]

All providers of intermediary services have to publish – at a minimum – a yearly report disclosing detailed information on: orders received from Member States' authorities; notices submitted through the notice-and-action mechanism; [67] content moderation engaged by providers' own initiative; complaints received through the complaint-handling system; and the use of automated means for content moderation.[68] For all these categories, providers must include details on type of allegedly infringing content, action taken, median time needed to react, as well as indicators of the accuracy rate and possible rate of error for automated systems.

Online platforms and search engines must also disclose information on the out-of-court dispute settlement mechanism, including outcomes and median time for completion. [69] Additionally,

---

[65] *Id.* art. 15.
[66] *Id.* art. 16.
[67] *Id.* The notice-and-action mechanism requires online platforms to establish clear and effective procedures for users or entities to report on illegal content. Once the platform receives a notice, it is considered to have "actual knowledge or awareness" and thus, not benefit anymore from the liability exemption provided in Article 15 of the e-Commerce Directive. The platform has an obligation to confirm the receipt to the sender and to notify the individual of its decision and indicate pathways for redress.
[68] *Id.* art. 15.
[69] *Id.* art. 21. The out-of-court dispute settlement mechanism provides a further option for appeal in case of complaints which cannot be resolved through the platform's internal complaint-handling mechanism. Individuals will be able to resort to an independent dispute resolution body to mediate – or issue binding decisions – on a

online platforms will have to communicate the average number of active users in each Member State every six months.[70]

The transparency reporting obligations for very large online platforms and search engines apply every six months. They must also include detailed information about the procedures and personnel involved in content moderation. More specifically, VLOPs and VLOSEs must disclose the "human resources that the provider [...] dedicates to content moderation [...] broken down by each applicable language of the Member States," their "qualifications and linguistic expertise," their "training and support," and the "indicators of accuracy and related information."[71] To mitigate privacy and breach of confidentiality matters, VLOPs and VLOSEs can classify the information contained in the transparency reports available to the public and only provide the complete reports to the Commission and the respective Digital Services Coordinator(s).[72]

B. Enforcement

In addition to the transparency requirements outlined above, the DSA establishes a complex enforcement process to guarantee oversight and ensure compliance. The provisions targeting enforcement and oversight are numerous. The sections below explain the powers awarded to the EU Commission and to Digital Services Coordinators. It is worth acknowledging that private individuals also enjoy private enforcement powers in the DSA. On one hand, traditional remedies through claims for damages and injunctive reliefs continue to be available at the national level,[73] on the other, individuals can rely on the internal complaint handling system and, if unsuccessful, resort to the mandatory out-of-court dispute settlement established in article 21. Nevertheless, this section focuses primarily on public enforcement through regulators.

The DSA clarifies that the EU Commission has exclusive "supervision, investigation, enforcement and monitoring" powers towards VLOPs and VLOSEs.[74] Such concentration of power in the

---

dispute with a platform. For a critical perspective on the out-of-court dispute settlement, *see* Jörg Wimmers, *The Out-of-Court Dispute Settlement Mechanism in the Digital Services Act: A Disservice to Its Own Goals*, 12 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L, 381 (2021).

[70] *Id.* art. 24.

[71] *Id.* art. 42.

[72] *Id.* art. 42(5).

[73] Miriam C. Buiten, *The Digital Services Act: From Intermediary Liability to Platform Regulation*, 12 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 361 (2021).

[74] *DSA*, Section 4.

Commission might be a reaction to the enforcement challenged faced by the GDPR.[75] In the GDPR, Data Protection Authorities (DPAs) of the providers' main establishment were tasked with overseeing and enforcing the regulation. Over time, the issues with such an approach became apparent. Understaffed and underfunded national DPAs have struggled to act against platforms and, even when action was taken, it failed to deter future infringements.[76] Instead, the DSA seems to follow the approach taken in competition law, where the EU Commission plays a central role in its enforcement.[77]

Article 66 clarifies that the EU Commission has investigatory powers to ensure compliance of VLOPs and VLOSEs with the DSA. First, the Commission has investigatory powers to initiate proceedings whenever the Commission suspects that platforms or search engines have committed an infringement.[78] Investigatory powers also include the ability to request information from VLOPs, VLOSEs and any person who might have information related to a suspected infringement, including independent auditors.[79] To be valid, information requests must specify "the legal basis and the purpose of the request," "what information is required and the set period within which it is to be provided."[80] Lastly, the Commission can also take interviews and statements,[81] and conduct inspections at the company's premises.[82]

The Commission also exercises a monitoring function, where it may take actions to ensure effective implementation of the DSA by VLOPs and VLOSEs. To meaningfully do so, the Commission can require

---

[75] Ilaria Buri & Joris van Hoboken, *The DSA Supervision and Enforcement Architecture*, DSA OBSERVATORY (June 24, 2022), https://dsa-observatory.eu/2022/06/24/the-dsa-supervision-and-enforcement-architecture/.
[76] Eliska Pirkova, *The EU Digital Services Act Won't Work without Strong Enforcement*, ACCESS NOW (Dec. 9, 2021), https://www.accessnow.org/eu-dsa-enforcement/.
[77] Experts have criticized such an approach, raising concerns over the future of separation of powers in the EU. As to VLOPs and VLOSEs, the Commission would be 1) policymaker (executive), 2) lawmaker (legislative), and 3) enforcer (judiciary). For a broader overview on the matter, *see* Suzanne Vergnolle, *Enforcement of the DSA and the DMA – What Did We Learn from the GDPR?*, *in* TO BREAK UP OR REGULATE BIG TECH? AVENUES TO CONSTRAIN PRIVATE POWER IN THE DSA/DMA PACKAGE (Heiko Richter, Marlene Straub & Erik Tuchtfeld eds., 2021), https://papers.ssrn.com/abstract=3932809 (last visited Apr. 3, 2023); Ilaria Buri, *A Regulator Caught Between Conflicting Policy Objectives Reflections on the European Commission's Role as DSA Enforcer*, *in* PUTTING THE DSA INTO PRACTICE: ENFORCEMENT, ACCESS TO JUSTICE, AND GLOBAL IMPLICATIONS (2023).
[78] *DSA*, art. 66.
[79] *Id.* art. 67.
[80] *Id.*
[81] *Id.* art. 68.
[82] *Id.* art. 69.

access to explanations to providers' databases and algorithms, and demand that providers keep a record of all necessary documents to assess implementation and compliance. The Commission may appoint independent external auditors and experts within its monitoring powers.[83]

As for the enforcement powers, the Commission can issue interim measures based on a prima facie finding of an infringement, in case of risk of serious damage to users.[84] Additionally, VLOPs or VLOSEs under an investigation can commit to specific steps toward ensuring compliance. The Commission has, then, the ability to make such commitments binding on the providers.[85] In case of established non-compliance, the Commission must provide the concerned VLOP or VLOSE with an opportunity to respond to the preliminary findings and take measures to ensure compliance.[86]

Suppose the measures taken as a response to the preliminary findings continue to be insufficient. In that case, the Commission can adopt a decision of non-compliance, ordering the concerned provider to take the necessary measures.[87] Suppose the measures continue to fail to ensure compliance within the established timeframe. In that case, the Commission can impose fines of up to 6% of the total worldwide turnover if the provider has intentionally or negligently infringed the DSA, failed to comply with the interim measures, or with a binding commitment.[88] The failure to provide information can result in fines of up to 1% of the total annual income or worldwide turnover,[89] or the imposition of periodic payments of up to 5% of the average daily income.[90]

Commentators have questioned the potential for penalties to deter infringements. Some have suggested that linking the exemption from liability to compliance would have been a more successful approach.[91] Indeed, previous compliance mechanisms that relied on the imposition of fines have proven unsuccessful as a means for deterrence.[92] At the same time, the penalty-based approach avoids over-blocking by platforms fearing liability. The DSA focuses on overseeing and

---

[83] *Id.* art. 72.
[84] *Id.* art. 70.
[85] *Id.* art. 71.
[86] *Id.* art. 73(2).
[87] *Id.* art. 73(1) & (3).
[88] *Id.* art. 74(1).
[89] *Id.* art. 74(2).
[90] *Id.* art. 76.
[91] Buiten, *supra* note 73.
[92] *See generally* Garry A. Gabison & Miriam C. Buiten, *Platform Liability in Copyright Enforcement*, 21 COLUM. SCI. & TECH. L. REV. 237 (2020).

controlling the processes that lead to content moderation, rather than the actual content.

Once the Commission has decided, an enhanced supervision system is to be established when the obligations are breached regarding the management of systemic risks.[93] The concerned platform or search engines must set up an action plan within the supervision system to terminate or remedy the infringement.[94]

Nevertheless, some uncertainties remain over the risk of overreaching powers by the Commission. It is important to stress that the EU Commission is a political body, representing the executive branch of the European Union. In the context of the DSA, the executive branch of the EU put forward the legislative initiative and moderated negotiations between the EU Council and the EU Parliament.[95]

The enforcement structure of the DSA concentrates a significant amount of power in the hands of the Commission towards oversight of specific platforms. Although such choice was supported by the need to avoid replicating the enforcement failures of the GDPR, it remains unclear how the conflicting policy interests of such a political body will be balanced. More specifically, the Commission's primary policy interests of safeguarding the internal market and incentivizing economic growth and innovation might be difficult to balance with the need to protect fundamental rights, and, as commentators have pointed out – might even influence the overall enforcement of the DSA.[96]

## III. WHAT DO TRANSPARENCY REPORTS TELL US?

This section provides a comparative analysis of transparency reporting practices among 19 VLOPs and VLOSEs before and after the implementation of the Digital Services Act. Prior to the DSA's enactment, transparency reporting was a voluntary practice, with notable variations in adoption and depth across platforms. Noteworthy examples include Google, Facebook, and Microsoft, which had already embraced transparency disclosures, especially regarding government orders and requests. In contrast, some platforms like AliExpress, Zalando, and Booking.com lacked transparency reports entirely.

---

[93] *DSA*, Chapter III, Section 5.
[94] *Id.* art. 75.
[95] Under the EU Ordinary Legislative Procedure, the EU commission mediates the discussions between representatives of the EU Parliament and the EU Council to agree on a "joint text." *See* Consolidated Version of the Treaty on the Functioning of the European Union, art. 294, O.J. (C 202/173) (2016), http://data.europa.eu/eli/treaty/tfeu_2016/art_294/oj/eng (last visited May 9, 2023).
[96] Buri, *supra* note 77.

The focus on VLOPs and VLOSEs designated by April 2023, prior to the DSA's transparency obligations taking effect, is deliberate.[97] This group encompasses a diverse array of platforms, including social media sites, search engines, and online marketplaces, leading to a wide range of reporting practices. This diversity is crucial for understanding the landscape of transparency reporting covered by the DSA.

In terms of methodology, the analysis was circumscribed to transparency aspects not entangled with other DSA mandates, such as the Notice-and-Action Mechanism or the Internal Complaint Handling Mechanism. [98] The evaluation framework for pre-DSA transparency reports was aligned with the obligations set forth in Articles 15(1)(a)-(c) and 42(2)(a)-(c) of the DSA, covering a comprehensive range of reporting criteria.

The assessment considered various forms of reporting, including both summary reports accessible on platforms' websites and detailed machine-readable formats, when available. This approach facilitated a nuanced comparison, focusing not only on the presence or absence of specific types of information (on a binary YES/NO basis) but also on the quality and granularity of the disclosed data. For the inaugural DSA compliance reports, the analysis was based on the official documents submitted by each platform to the EU Commission. This provided a direct insight into how platforms have adjusted their transparency reporting practices in response to the new regulatory framework established by the DSA.

The following analysis is structured into three distinct subsections to delve deeper into specific areas of transparency reporting. The first subsection scrutinizes how platforms reported on government takedown orders and information requests, shedding light on the interaction between online platforms and governmental entities in regulating online content. The second subsection explores reporting on voluntary content moderation practices, focusing on the autonomous efforts by platforms to monitor and manage the content they host, beyond legal mandates. The third and final subsection examines reporting on human resources involved in content moderation, offering insights into the scale and nature of the workforce dedicated to maintaining online safety and compliance with content policies. This tripartite structure enables a comprehensive understanding of the multifaceted approaches to

---

[97] The analysis excludes Aylo Freesites Ltd. (Pornhub), Technius Ltd. (Stripchat), and WebGroup Czech Republic (XVideos), which were designated in December 2023. *See* European Commission, *supra* note 64.

[98] *DSA,* arts. 16 & 20.

transparency in the digital sphere, and how they were impacted by the entry into force of the DSA.

### A. Government Requests

#### 1. Pre-DSA Practices

The move towards transparency in handling government takedown and information requests predates the DSA. The initiation of transparency reporting practices can be traced back to the early 2010s, with companies like Google and Twitter pioneering the effort as a form of recognition of their role in safeguarding democratic values.[99] As debates around privacy, government surveillance, and freedom of expression intensified globally, leading platforms began to recognize the need to establish trust with their user base and the broader public.[100]

Google, for instance, launched its first Transparency Report in 2010, setting a precedent for the industry.[101] This initial report was a landmark event, signaling a shift towards greater openness about government requests for user information and demands for content removal.[102] Google's move was both a response to growing public concern about online privacy and an attempt to position itself as a transparent and accountable entity in the face of increasing government requests. Twitter followed suit, releasing its own transparency reports that detailed government requests for user information and content takedowns.[103] These early reports from Google and Twitter not only showcased the companies' apparent commitments to user rights and transparency but also shed light on the scale of government surveillance and censorship efforts worldwide.

The practice of transparency reporting quickly gained traction among other major platforms, particularly within the realms of social media and search engines, gradually evolving into an industry

---

[99] Suzor et al., *supra* note 46.
[100] Aleksandra Urman & Mykola Makhortykh, *How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms*, 47 TELECOMMS. POL'Y 102477 (2023).
[101] David Drummond, *Tools to Visualize Access to Information*, OFFICIAL GOOGLE BLOG (Sept. 20, 2010), https://googleblog.blogspot.com/2010/09/tools-to-visualize-access-to.html (last visited Feb 26, 2024).
[102] Claire Cain Miller, *Google Reports on Government Requests and Censorship*, N.Y. TIMES BITS BLOG (Sept. 21, 2010), https://archive.nytimes.com/bits.blogs.nytimes.com/2010/09/21/google-reports-on-government-requests-and-censorship/
[103] Jeremy Kessel, *Twitter Transparency Report*, X BLOG (July 2, 2012), https://blog.x.com/en_us/a/2012/twitter-transparency-report (last visited Feb 26, 2024).

standard.[104] As of April 2023, most platforms considered in this analysis had introduced a reporting system for government orders – the only exceptions being AliExpress, Booking.Com, and Zalando.

However, upon closer look at the platforms that did engage in transparency reporting of government takedown requests, it is possible to notice that there has been an uneven adoption of the practice. While most companies provide country-by-country information on how many requests were received and how many were fulfilled, further granularity – such as the alleged legal basis for the takedown request – was usually lacking besides a few exceptions. For instance, Apple's App Store transparency reports stood out for their depth, breaking down requests by country, number of apps per request, government entity, and the specific legal basis invoked.[105] This level of granularity was not universal, with platforms like Snapchat offering a more cursory view, focusing on aggregate figures that offered limited insight into the specifics of government requests.[106]

For what concerns government information requests, the granularity disclosed was slightly more uniform, with most platforms presenting data by country and distinguishing between emergency requests, other information requests, and providing percentages of requests where some data was produced. Nevertheless, information on the requesting government entity and the specific legal basis involved continued to be lacking in the majority of reports.[107] This approach to transparency reporting has been criticized for being a "communication trick" rather than a way to promote the principle of transparency.[108] Without insights on what information is produced or what requests are

---

[104] Kosta & Brewczyńska, *supra* note 14, at 5–6. (also stressing the role played by the 2013 Snowden revelations in the industry-wide move towards transparency over government requests).

[105] *2022 App Store Transparency Report*, APPLE (2023), https://www.apple.com/legal/more-resources/docs/2022-App-Store-Transparency-Report.pdf (last visited Feb 26, 2024). ("CSV file, 'app_takedown_platform_policy_violation_requests.csv'")

[106] *Snapchat Transparency Report*, SNAPCHAT PRIV. SAFETY, AND POL'Y HUB (2023), https://values.snap.com/privacy/transparency (last visited Feb 26, 2024).

[107] *See id.*; *Government Requests for User Data*, META TRANSPARENCY CTR. (2023), https://transparency.fb.com/reports/government-data-requests/ (last visited Feb 27, 2024); *Transparency Report*, PINTEREST POL'Y (2023), https://policy.pinterest.com/en/transparency-report (last visited Feb 27, 2024); *Information Requests Report*, TIKTOK (Nov. 10, 2023), https://www.tiktok.com/transparency/en/information-requests-2023-1/ (last visited Feb 27, 2024); X, *Information Requests Transparency Report H2 2021*, X TRANSPARENCY CENTER (2022), https://transparency.x.com/en/reports/information-requests.html (last visited Feb 27, 2024).

[108] Kosta & Brewczyńska, *supra* note 14, at 2, 10. (explaining that transparency reports usually include "statements on the role of these reports, as perceived by a company, and the targeted audience" signaling to the public the rationale behind this voluntary reporting initiative).

not reflected in the report, receivers of this information have no means of acting upon it.[109]

Detailed reporting, such as categorizing requests based on the nature of the alleged infringement, is not merely a transparency exercise. [110] It plays a pivotal role in public discourse, enabling stakeholders to discern patterns of government oversight and potential overreach. Such details help contextualize the legal and social challenges confronting digital platforms and their users, ranging from free expression to privacy rights. Yet, the content of these reports is often inconsistent, presented in an aggregate manner, completely lacking uniformity in structure and content, and the reporting manner further complicates the ability to compare and evaluate them. These varying approaches reflect differing corporate policies, legal advisories, and perhaps strategic considerations about how much to disclose. Ultimately, while platforms tend to promote themselves as enablers of transparency, the information provided in voluntary transparency reports leaves the actual state of government access and takedown requests opaque.[111]

2. First DSA Reports

The advent of the DSA marked a significant shift in the regulatory landscape, imposing specific requirements for transparency obligations on government takedown and information requests.[112] At first sight, it appears that the obligations under Article 15 did in fact force platforms to provide more detailed information on government requests. At the same time, however, the initial reports under the DSA revealed what seems to be a contraction in the breadth of reporting. Notably, neither Google's platforms nor Meta's reported on government removal requests – claiming to have not received a single request - a surprising development given the companies' history of transparency. Nevertheless, the observed reduction in reporting on government takedown and information requests in the post-DSA era may not solely be indicative of a decrease in transparency or a reluctance by platforms to disclose

---

[109] *Id.* at 1.

[110] This is aligned with the initial driver of platform transparency in the early 2010s, when platforms started to disclose government requests received, portraying themselves as protectors of democratic values. The granularity of the information disclosed, however, varied. Only a few platforms – Apple, Google, LinkedIn, Pinterest, TikTok, and X – provided specific information over the category of alleged infringement.

[111] *See also* Kosta & Brewczyńska, *supra* note 14 (comparing government request transparency reports).

[112] *DSA,* art. 15(1)(a).

government interactions.

An alternative explanation for this trend could lie in the evolving dynamics between platforms and governments, particularly with the increasing reliance on the notice-and-action mechanism stipulated by the DSA.[113] This mechanism allows individuals and entities to notify platforms of potentially illegal content, prompting a review and, if deemed necessary, subsequent removal of such content. The shift towards this more decentralized approach to content moderation could mean that governments are opting to use the notice-and-action mechanism as a more immediate and less formal avenue for content regulation, as opposed to the traditional direct government requests for takedowns or information.[114]

This explanation, however, raises several intricate questions about the interplay between government entities and platforms in content moderation. When government entities act under Article 16 instead of Articles 9 and 10, the distinction between their role and direct government orders for content takedowns and information requests becomes blurred. One pivotal concern is the equivalence of actions taken by government entities as notifiers and those taken under direct orders pursuant to articles 9 and 10. The latter comes with a set of safeguards designed to ensure due process and protect the rights of users. These safeguards include the requirements for clear, specific orders, and the duty to notify the user when effect is given to an order, among others.[115] When the government bypasses these formalities by using the notice-and-action mechanism, it potentially circumvents the protective measures embedded in the DSA, raising questions about the legitimacy and accountability of such actions.

Among the platforms that did report on government takedowns, a continued variance in the level of detail is evident. Notably, not even half of the platforms provided a breakdown of requests by specific categories

---

[113] Under the notice-and-action mechanism, individuals or entities can notify platforms of the presence of allegedly illegal content. Notices are considered to give rise to actual knowledge or awareness for liability purposes. *See DSA* art. 16 & art. 6.

[114] This transition towards the notice-and-action mechanism might not necessarily translate to platforms disclosing less information about their interactions with government entities. Rather, it could signify a shift in the nature of these interactions, moving away from formal requests that would be captured in traditional transparency reports to more indirect forms of government influence on platform content policies and enforcement actions. Since the analysis presented here does not cover transparency over the notice-and-action mechanism, the apparent reduction in reporting on direct government requests does not provide a complete picture of the full spectrum of government-platform interactions in the content moderation ecosystem.

[115] *DSA,* art. 9.

of alleged illegal content or behavior, opting instead for broad, self-defined categories such as "Provides or facilitates an illegal service" as reported by Apple App Store, "Unsafe and/or Illegal Products" by Booking.Com, and "Illegal Hate Speech" in other instances. The absence of references to specific laws in these reports diminishes the utility of this transparency mechanism, particularly if one of the objectives is to hold governments accountable. For example, the term 'illegal hate speech' can have varied legal definitions and bases in different Member States (MS), making it imperative to cite specific legal provisions to enhance the clarity and effectiveness of these reports.[116]

When it comes to information requests, only a select few companies provided detailed insights into the legal bases cited by requesting authorities, with TikTok delineating 27 categories, [117] X identifying 14,[118] and Facebook and Instagram detailing 22 categories of alleged illegal activity.[119] However, once more, data presentation varies hindering comparability. For example, the manner in which Facebook and Instagram present this data – in two separate tables sorted by MS and by category of illegal content – complicates the task of pinpointing the specific categories of requests issued by each MS.[120] This level of detail is crucial, particularly for researchers and civil society organizations aiming to understand the legal grounds on which governments engage with platforms.

The DSA also mandates platforms to report median times for both acknowledging receipt of requests and handling them. The reports reveal that acknowledgment of receipt is typically automated upon submission for most platforms, with exceptions including AliExpress, Amazon, and the App Store. The reported median handling times exhibit considerable variability, with some platforms measuring in hours and others in days, complicating comparative analysis across different platforms. This variability is further exacerbated by orders that span multiple accounts

---

[116] Although the DSA requires companies to takedown illegal content, the definition of what constitutes illegal content is left to the Member States. *DSA*, art. 6.
[117] *TikTok's DSA Transparency Report 2023*, TIKTOK (Sept. 2023), https://www.tiktok.com/transparency/en/dsa-transparency/ (last visited Feb 26, 2024) at 18-21.
[118] *DSA Transparency Report 2023*, X (2023), https://transparency.x.com/dsa-transparency-report-2023.html (last visited Feb 26, 2024).
[119] *Regulation (EU) 2022/2065 Digital Services Act: Transparency Report for Facebook*, META TRANSPARENCY CTR. (2023), https://transparency.fb.com/sr/dsa-transparency-report-oct2023-facebook/ (last visited Feb 26, 2024), at 5-6; *Regulation (EU) 2022/2065 Digital Services Act:Transparency Report for Instagram*, META TRANSPARENCY CTR. (2023), https://transparency.fb.com/sr/dsa-transparency-report-oct2023-instagram/ (last visited Feb 26, 2024), at 5-6.
[120] *Id.*

or content pieces, where it remains unclear how response times are aggregated to calculate the median handling time. All reports lack contextual information on how median handling times are computed, especially in instances of partial compliance with takedown orders, such as when a platform complies with some parts of an order but not others.

The current DSA-mandated reporting framework, with its emphasis on quantifying government orders received and complied with, simplifies the complex array of responses platforms may deploy when confronted with government requests. This binary reporting system does not capture the nuanced reality of partial compliance, leaving ambiguities in how such cases are categorized and understood. The necessity for granularity in reporting on government requests transcends the realm of platform transparency, extending into the domain of empowering civil society to hold governments accountable.[121] Enhanced specificity and detail in transparency reports are indispensable for fostering a more informed and accountable digital public sphere.

### B. Own-Initiative Content Moderation

#### 1. Pre-DSA Practices

In the pre-DSA landscape, the approach to transparency reporting on own-initiative content moderation practices by platforms was marked by significant disparity among each other, both in terms of frequency and granularity. A majority of platforms published at least an annual report

---

[121] *See* Frederik Stjernfelt & Anne Mette Lauritzen, *The Role of Civil Society*, *in* YOUR POST HAS BEEN REMOVED: TECH GIANTS AND FREEDOM OF SPEECH 241 (Frederik Stjernfelt & Anne Mette Lauritzen eds., 2020), https://doi.org/10.1007/978-3-030-25968-6_17 (last visited Feb 19, 2024). (on the role of civil society in holding government accountable for its interplay with online platforms). *See also The Santa Clara Principles on Transparency and Accountability in Content Moderation*, SANTA CLARA PRINCIPLES, https://santaclaraprinciples.org/images/santa-clara-OG.png (last visited Feb 19, 2024). The Santa Clara Principles were formulated with significant input from civil society organizations, underscoring the critical role these groups play in advocating for transparency and fairness in online content moderation. Conceived during a 2018 gathering in Santa Clara, California, these principles reflect a collaborative effort between academic experts, civil liberties organizations, and other stakeholders concerned with digital rights. Civil society's involvement was pivotal in highlighting the necessity for greater accountability and user rights protection in content moderation processes. The principles advocate for enhanced transparency in content removals and account suspensions, the right to appeal moderation decisions, and the proportionality of moderation actions. They emphasize the importance of detailed reporting on moderation practices, aiming to ensure that platform policies are implemented in a manner that respects free expression and fair treatment of users. Through these principles, civil society continues to influence the discourse on digital governance, promoting standards that safeguard user rights in the evolving landscape of online platforms.

detailing their internal content moderation activities, while others published quarterly reports.[122]

Voluntary reports provided aggregated data informing users on how each platform enforced its content policy. The reports analyzed typically included data categorized by the type of policy violated, offering insights into the specific nature of content being moderated, with greater emphasis on terrorist content and child sexual abuse material (CSAM). However, in most cases, an interesting discrepancy emerged in the alignment between the categorizations used in these transparency reports and those outlined in the platforms' own community guidelines or policies. For instance, Facebook's Community Guidelines Enforcement Report (CGER) was structured around 11 policy areas, despite its Community Standards encompassing a broader set of 24 rules. This mismatch raises questions about the coherence and comprehensiveness of transparency reporting in fully reflecting the platforms' moderation policies.

The geographical granularity of reported data further varied across platforms. While some, like the Apple App Store,[123] provided detailed breakdowns of moderation decisions by country, others, including Facebook, did not offer country-specific data.[124] This lack of geographical specificity in reporting can obscure the regional nuances of content moderation practices and their impact on different user communities.

Another dimension where transparency reporting practices varied significantly among platforms was in the disclosure of content restrictions mandated by law. Not all platforms elected to incorporate this crucial information within their general transparency reports. Instead, there was a tendency for platforms to publish separate country-specific reports to address legal content restrictions.[125] This approach, while providing localized insights, complicates efforts to conduct cross-comparisons and understand the global impact of legal requirements on

---

[122] Notable exceptions existed, including AliExpress, Amazon, Booking.com, Wikipedia, and Zalando.

[123] Apple, *supra* note 105, at 1 ("Apps Removed from the App Store Subject to Government Takedown Demands").

[124] *Community Standards Enforcement Report Q2 2023*, META TRANSPARENCY CTR., https://transparency.fb.com/reports/community-standards-enforcement/ (last visited Feb. 26, 2024).

[125] *See id.*; *YouTube Community Guidelines Enforcement*, GOOGLE TRANSPARENCY REP. (2023), https://transparencyreport.google.com/youtube-policy/removals?hl=en (last visited Feb. 27, 2024); *Community Guidelines Enforcement Report April 1, 2023-June 30, 2023*, TIKTOK (Oct. 12, 2023), https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2023-2/ (last visited Feb. 27, 2024).

content moderation practices. The fragmentation of reporting into country-specific documents can obscure the broader patterns of legal influence on platform operations, making it challenging to assess the consistency and comprehensiveness of platforms' responses to legal mandates across different jurisdictions.

The role of automation in content moderation was broadly acknowledged by platforms, with many citing reliance on automated systems for screening content. Yet, detailed disclosures about the functioning of these automated moderation systems were less common and, when provided, often focused narrowly on specific areas like the detection of child sexual abuse material (CSAM) or copyright violations.[126] The accuracy and error rates of these automated tools are seldom mentioned in the reports, leaving a significant gap in understanding their reliability and potential for incorrect decisions.

Furthermore, platforms demonstrate a marked reluctance to disclose detailed information about the interplay between automated tools and human reviewers, which would be critical for understanding content moderation's inner workings. Detailed insights into how platforms balance automation with human review could reveal the safeguards against automation biases and errors, and how human reviewers are trained to address the complexities that automated systems might overlook. This opacity, however, leaves a significant void in public knowledge about how decisions are made, with most insights coming from leaked documents and whistleblower testimonies.[127]

It was also frequently difficult to ascertain from the reports what portion of the disclosed content moderation decisions were made solely by automated means or by human reviewers being supported by automated means. This distinction is crucial as the implications for

---

[126] *See Digital Safety Content Report*, MICROSOFT CORP. SOC. RESP. (2023), https://www.microsoft.com/en-us/corporate-responsibility/digital-safety-content-report (last visited Feb. 27, 2024); Meta, *supra* note 124; YouTube, *supra* note 125, at 6–7.

[127] These unofficial sources have been instrumental in shedding light on the realities of content moderation, revealing the complexities and challenges inherent in balancing automated screening with human judgement. However, the reliance on such leaks and whistleblowers to obtain information underscores the need for a more forthright approach. Without it, the opacity on this matter will continue to raise concerns about the potential for errors and biases within moderation systems. See e.g., Jeff Horwitz, *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.*, WSJ, Sep. 13, 2021, https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353; Deepa Seetharaman, Jeff Horwitz & Justin Scheck, *Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts.*, WALL STREET JOURNAL, Oct. 17, 2021, https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184.

content moderation accuracy, fairness, and user trust can vary significantly between the two. Furthermore, information on automated moderation practices was often relegated to platforms' blog posts, separate from the formal transparency reports, which may dilute the visibility and accountability of these practices. [128] This lack of transparency matters because it impedes the ability of observers to evaluate the efficacy of automated systems and assess the potential for systemic biases, and leaves unanswered questions about the overall governance and accountability mechanisms in place to moderate content.

Regarding the measures taken to limit the availability, visibility, and accessibility of content, platforms' disclosures were generally vague. Most platforms provided data on actions taken against content and accounts, such as content removals, account suspensions, and deletions. However, more nuanced interventions like geo-blocking, de-ranking, or restrictions on account interactions were seldom disclosed. Similarly, the practice of labeling content, was not widely reported, leaving a gap in understanding the full spectrum of moderation actions employed by platforms.

The analysis of voluntary transparency reports underscores a heterogeneous field of practices among platforms. The variability in reporting frequencies [129], the alignment between transparency report categories and content policies, the level of geographical detail, the inclusion of legally mandated restrictions, the detailing of automated moderation practices, and the disclosure of specific moderation measures collectively illustrate a complex and multifaceted transparency landscape within the digital platform ecosystem.

### 2. First DSA Reports

The DSA introduced specific reporting requirements on platforms' own moderation practices. The first DSA reports are characterized by significant advancements in the depth and structure of

---

[128] *See* Microsoft, *supra* note 126, (FAQ); *How Meta Enforces Its Policies*, META TRANSPARENCY CTR. (2023), https://transparency.fb.com/enforcement/ (last visited Feb. 27, 2024).

[129] The reporting frequency varied significantly among platforms, with some publishing quarterly reports (Facebook, Instagram, and TikTok), and some publishing biannual reports (LinkedIn, Pinterest, Snapchat, Twitter). Twitter's last report before the DSA Report refers to the second half of 2021. *See* Twitter Rules Enforcement Report Jul-Dec 2021, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec.

transparency reporting,[130] though it also reveals that several pre-existing challenges persist, underscoring the complexity of achieving comprehensive transparency in content moderation.

One of the key developments in the post-DSA landscape is the enhanced categorization of content moderation decisions. Unlike the pre-DSA reports, where the disclosure of content moderation practices varied widely in terms of granularity and detail, the post-DSA reports show a marked improvement with platforms categorizing their moderation decisions by the type of infringement, detection method, and type of restriction applied. This structured approach reflects a significant stride towards greater transparency. However, the diversity in how platforms approached the task of categorization complicates direct comparisons, a challenge reminiscent of the pre-DSA era where inconsistencies in reporting practices were common.

Post-DSA, an encouraging trend is the widespread adoption of categorizing moderation decisions by the type of infringement, a notable advancement from the pre-DSA era where such detailed categorization was less common. Platforms, with the exceptions of Amazon and Zalando, have embraced this approach, enhancing the transparency of their moderation activities. Amazon's reports, interestingly, opted for categorization by product type rather than the nature of the infringement, while Zalando reported no own-moderation initiatives. This categorization is crucial for understanding the scope of content moderated but introduces complexities due to the lack of standardization in how violations are defined. Even within entities under the same corporate umbrella, the inconsistency is evident with significant disparity in how categories are defined.[131]

The geographical breakdown of content moderation, a critical aspect for understanding the regional impact of these practices, remains largely unaddressed in the first DSA reports. Platforms like TikTok, Google, and Instagram have not provided data broken down by Member State, continuing a trend from the pre-DSA era where such specificity was rare. This omission obscures the localized nuances of moderation

---

[130] Including the issuing of the first-ever reports on own initiatives by AliExpress, Amazon, Booking.Com, Wikipedia, and Zalando.

[131] Facebook and Instagram reported 13 and 12 categories respectively, illustrating the variability in how violations are defined and categorized. Meta, *supra* note 119, at 10–11. Google Search lists 17 categories, Google Play lists 14, Google Shopping lists 15, Google Maps lists 11, and YouTube lists 11. *EU Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report*, GOOGLE (Oct. 27, 2023), https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf (last visited Feb. 27, 2024).

practices, making it difficult to assess the platforms' responsiveness to regional content norms and legal requirements.

The requirement to break down content moderation decisions by detection method was interpreted in different ways by platforms, potentially signaling a lack of specificity in the DSA provision. While most platforms distinguished between total decisions and decisions made by solely automated means,[132] Pinterest stood out categorizing each moderation category by detection method – manual, automated, hybrid – offering a more granular view. Contrariwise, Google's platforms provided a less nuanced differentiation by distinguishing between "self-detection," "external detection," and "unknown detection," making it impossible to assess the impact of automated detection tools.[133] Interestingly, such breakdown was completely absent in Amazon, Bing[134], and Snapchat's reports.

The DSA also mandates platforms to report on "indicators of the accuracy and the possible rate of error of the automated means used" in content moderation.[135] Although this requirement aimed to enhance transparency and accountability, it also introduced complexities due to the absence of explicit definitions for key terms within the DSA itself. The approaches to defining and reporting on the accuracy and error rates of automated moderation systems vary significantly across platforms, reflecting the diverse methodologies and interpretations employed. TikTok, for instance, provided a clear definition of these metrics, with the "error rate" reflecting the proportion of videos and ads reinstated after an appeal, and the "accuracy rate" representing the proportion of content

---

[132] *Booking.Com Digital Services Act*, BOOKING.COM 4–5 (Oct. 27, 2023), https://r-xx.bstatic.com/data/mobile/dsa_transparency_report_bf3fdc24.pdf (last visited Feb. 27, 2024); TikTok, *supra* note 117, at 9; X, *supra* note 118, at 2; *AliExpress DSA Biannual Transparency Report*, ALIEXPRESS https://www.aliexpress.com/p/transparencycenter/reports.html (last visited Feb. 27, 2024); Apple App Store, *DSA Transparency Report*, APPLE LEGAL (Oct. 2023), https://www.apple.com/legal/dsa/transparency/eu/app-store/2310/ (last visited Feb. 27, 2024).

[133] Google platforms distinguish between self-detection (which comprehends employees, algorithms, or contractors flagging content), external detection (user policy flags or legal complaints), and unknown detection (system limits). Google, *EU Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report*, (2023), https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2023-8-28_2023-9-10_en_v1.pdf.

[134] Bing did not provide a breakdown by type of restrictions imposed, instead it only provided the total amount of actions taken. *Microsoft Bing Transparency Report (Regulation (EU) 2022/2065),* https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1dOoh (October 2023).

[135] *DSA,* 15(1)(e).

that remains un-reinstated post-appeal.[136] X (formerly Twitter) adopted a somewhat similar approach by reporting "overturn rates."[137] This approach, however, included all types of suspensions, not solely those initiated by automated means, potentially blurring the focus on automated moderation's specific accuracy and error rates. Google distinguished itself by utilizing a sampling method to assess the accuracy of its automated moderation, where human evaluation of a random sample of user contributions serves as the benchmark for assessing the automated system's decisions. While providing depth, this unique methodology complicates direct comparisons with other platforms due to its distinct approach.[138] Contrastingly, many platforms opted for qualitative descriptions over quantitative data, providing narratives that outline the functioning of their automated moderation systems but lacking the specificity and clarity afforded by numerical metrics. This qualitative approach, seen in platforms like Facebook, Instagram, and Snapchat, while informative, does not afford the same level of precision and transparency as quantitative metrics.

Nevertheless, the information disclosed on these criteria remains inconsistent and often vague. None of the platforms accurately described how the accuracy of automated tools is measured, with the majority of platforms relying on the percentage of content reinstated upon appeal to measure the error rate.[139] Despite these disclosures, the overall lack of detailed information on the accuracy and error rates of automated tools, coupled with the absence of clear explanations regarding the interplay between automated and human moderation, remains a significant challenge. This lack of clarity hinders the ability to fully assess the reliability, fairness, and potential biases inherent in automated moderation systems.

While the post-DSA era has seen advancements in the transparency of content moderation practices, with more detailed categorization of moderation decisions and enhanced disclosures on

---

[136] TikTok, *supra* note 117, at 3–4. This method offers a tangible measure of TikTok's automated moderation system's reliability.

[137] Which, although not explicitly labeled as "error rate," serve a similar purpose by indicating the proportion of enforcement actions overturned upon appeal.

[138] Google, *supra* note 132, at 24–26. ("[A]ccuracy is computed based on human evaluation of a random sample of all user contributions, across data types and content types (e.g., reviews, media, facts, etc.) between 1 March 2023 and 31 August 2023. The accuracy for that slice is then defined as the percentage of correct decisions made by the automated system, assuming the human evaluation is the ground truth.")

[139] It is interesting to note that Google's own platforms assess accuracy in different ways. Google Search and Google Play use "precision rate" only (whereas accuracy is used for Maps). YouTube only provides the Violative View Rate (estimate of the proportion of video views that violate policy). *Id.* at 28.

detection methods, significant challenges persist. The lack of a standardized approach to categorizing violations, the absence of geographical breakdowns in reporting, and the incomplete disclosure regarding the effectiveness of automated moderation tools continue to complicate the comparative analysis of platforms' content moderation practices. Addressing these issues is crucial for advancing transparency and accountability in the digital ecosystem.

### C. Human Resources Involved in Content Moderation

#### 1. Pre-DSA Practices

The discourse on the human resources dedicated to content moderation within digital platforms has emerged as a critical area of inquiry. The reticence of platforms to unveil the intricacies of their content moderation infrastructure, including the specifics of moderation teams, their linguistic competencies, qualifications, and the nature of their training, have historically fueled speculation regarding the impartiality and contextual adequacy of moderation decisions.

Human reviewers play a crucial role in the complex task of deciding whether user-uploaded content should be enabled on a specific platform, requiring a deep understanding of the platform's policies and how to enforce them effectively.[140] To be able to moderate content correctly, reviewers are also expected to be proficient in the relevant languages, understand the applicable content laws, and have received comprehensive training on how to interpret and apply the platform's policies. Additionally, considering the nature of content they are often exposed to, human reviewers should receive sufficient psychological support both during and after employment, as well as adequate pay.

Nevertheless, the picture portrayed by whistleblowers and leaked documents appears to be rather different. Often, professionals are contracted in regions where labor costs are lower than those in the primary user bases and headquarters of the platforms, and where labor laws tend to be less demanding of the employer. Furthermore, the transparency regarding the internal content review guidelines, working conditions, and support systems has unsurprisingly been lacking. Platforms have traditionally shielded these details under the guise of proprietary information, arguing that disclosing them could enable malicious actors to exploit the system. Prior to the DSA, the available

[140] Sarah T. Roberts, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

insights into these practices were predominantly sourced unofficially, such as through leaked documents and legal proceedings, rather than from the platforms themselves.[141] Notably, judicial cases brought against platforms, including the ongoing case in Kenya where former content moderators are suing Facebook over trauma suffered due to the nature of their employment, have shed light on the working conditions, challenges, and the often-overlooked human aspect of content moderation.[142] Although limited, these cases already highlight not only the psychological toll on moderators but also raise questions about the adequacy of their training, support, and the overall sustainability of the content moderation ecosystem.

The analysis of reports from this period reveals a stark absence of substantial information on the composition of content moderation teams, their language skills, or the specifics of their training regimes. At best, some platforms provided vague references to the number of individuals involved in moderation, hinting at the balance between automated detection and human review without providing concrete details.

### 2. First DSA Reports

The DSA introduced a mandate for platforms to disclose and qualitatively describe the human resources involved in content moderation, representing a significant stride towards enhancing transparency. Initial reports under the DSA have seen platforms disclosing the number of content moderators per EU official language and providing some insight into their qualifications and training.

---

[141] Martha Dark, *Revealed: Accenture Forces Its Facebook Moderators to Sign a Form Acknowledging that the Work Can Lead to PTSD*, FOXGLOVE (Jan. 28, 2020), https://www.foxglove.org.uk/2020/01/28/revealed-accenture-forces-its-facebook-moderators-to-sign-a-form-acknowledging-that-the-work-can-lead-to-ptsd/ (last visited Feb. 27, 2024); Vittoria Elliott, *A Leaked Memo Shows TikTok Knows It Has a Labor Problem*, WIRED (July 21, 2023), https://www.wired.com/story/tiktok-leaked-documents/; Alex Hern, *Is the End Nigh for End-to-End Encryption?*, THE GUARDIAN (Apr. 2, 2022), https://www.theguardian.com/commentisfree/2022/apr/02/is-the-end-nigh-for-end-to-end-for-encryption-whatsapp.

[142] Rodney Muhumuza & Amanda Seitz, *Facebook Sued in Kenya over Work Conditions for Moderators*, AP NEWS (May 10, 2022), https://apnews.com/article/business-lawsuits-africa-nairobi-uganda-a93d3e60bcebee3124b2d2b168c652dc; *See also* The Associated Press, *Facebook to Pay Moderators $52M for Psychological Damages*, AP NEWS (May 12, 2020), https://apnews.com/general-news-faa5df03e40f6b9736225e49d8ceaf19 (In 2020 Facebook agreed to pay former U.S.-based moderators $52 million after "repeated exposure to graphic material such as child sexual abuse, beheadings, terrorism, animal cruelty and other disturbing images.")

However, the level of detail remains insufficient for a comprehensive understanding of content moderation practices. While disclosures now include the linguistic capabilities of moderation teams, further clarifications regarding the moderators' qualifications, location, or working conditions are absent, limiting the depth of insight into the global moderation infrastructure. Platforms' responses to the DSA's requirements on disclosing moderator qualifications have varied widely, with some listing academic degrees and others describing moderators' roles within the company or their tenure.[143]

Regarding the training and support extended to content moderation teams, the disclosed information varies widely across platforms, potentially due to its qualitative nature. While a majority of platforms vaguely mention that reviewers undergo both initial and ongoing training, no platform has provided concrete examples of the training methodologies employed. Some platforms have acknowledged that the training regimen differs based on the focus area, with teams handling particularly harmful or potentially illegal content receiving additional training. Yet, the details remain vague and lack the granularity needed for meaningful analysis.

For example, TikTok's report highlighted "tools and features" designed to mitigate exposure to graphic content, including gray scaling and blurring, and mentioned training for managers to recognize when team members might need additional well-being support.[144] Google went a step further by providing post-exit mental health support, including counseling services, for employees exposed to sensitive content.[145] In contrast, Facebook[146] and Instagram[147] provided broad statements about the support systems in place for human reviewers, with some assertions about support requirements in vendor contracts, including aspects like pay, benefits, work environment, and psychological support.

This variance in the depth and specificity of information provided by platforms underscores the need for a more structured approach to transparency reporting of human resources involved in content moderation. Providing greater clarity on the terminology used and the

---

[143] For instance, Zalando's report provided a simple list of degrees, such as "Bachelor in Tourism," and general professional descriptions for its 20 part-time content moderators. *Transparency Reporting on Content Moderation*, ZALANDO (2023), https://mosaic02.ztat.net/cnt/contentful-apps/uploads/a74cdebf-cfc7-46dd-8853-13afed1e41aa.pdf.

[144] TikTok, *supra* note 117, at 5.

[145] Google, *supra* note 132, at 30.

[146] Meta, *supra* note 119, at 18–19. (Facebook).

[147] Meta, *supra* note 119, at 17. (Instagram).

expected information to be reported could encourage platforms to share more detailed and actionable insights, thereby enhancing the overall transparency of content moderation practices and contributing to a more informed and accountable digital ecosystem.

## IV. HOW TO ENSURE ACTIONABLE TRANSPARENCY MANDATES

The DSA had the ambitious goal of becoming what has been called a "transparency machine" – forcing platforms to unveil the opaque practices of content moderation that have long been shielded from public scrutiny. This ambition was to position the DSA as a cornerstone in the architecture of platform governance, promoting an unprecedented level of openness. When the transparency provisions of the DSA were first announced, many scholars and civil society representatives were optimistic, perceiving the Act as a conduit to obtain insights into platform operations, enforcement of internal policies, content removal patterns, and the role of moderation teams.

However, the first transparency reports fell short of these high expectations. The absence of a standardized reporting framework led to documents that were challenging to interpret, undermining the DSA's objective of fostering accessible transparency. [148]

As global regulators look at the DSA as a template for their own digital governance frameworks, it is crucial to assess these shortcomings. Enhancing standardization and verifiability emerges as pivotal to ensuring transparency mandates are not only met but are genuinely actionable.

### A. Standardization: The Keystone of Actionable Transparency

To effectively clear the haze around content moderation practices among social media companies and harness the full potential of transparency mandates, a comprehensive approach to standardization in transparency reporting is needed. Only through a unified and standardized reporting framework can transparency reports become meaningful and actionable, enabling cross-comparison, identifying trends, and ensuring compliance with regulatory frameworks like the DSA.[149]

The challenge of standardization, or rather the lack thereof, was a

[148] The DSA stresses that information needs to be easily understandable and accessible. *DSA*, art. 15(1).
[149] *See generally* Fung, Graham &Weil, *supra* note 12, at 43.

common thread running through all transparency requirements under the DSA. The diversity in metrics and granularity used by different platforms made it exceedingly difficult to compare and evaluate the effectiveness of content moderation processes across the board. This issue was further compounded by the misalignment in the interpretation of specific requirements of Articles 15 and 42, leading to a distorted view of the available information and, consequently, a failure to leverage transparency as a vector for compliance.

The need for standardization in transparency reporting is not unique to the digital sphere. The Global Reporting Initiative (GRI), for instance, has demonstrated the efficacy of standardization in enhancing transparency around corporate, social, and environmental responsibility. [150] Similarly, the International Financial Reporting Standards have facilitated the comparability of financial disclosures globally, highlighting the value of standardized reporting practices. [151] Drawing parallels from these examples, it is evident that standardization in the context of content moderation transparency is not just beneficial but necessary.

The push for standardized transparency reporting in content moderation has gained traction over the last few years. In 2020, the EU Commission itself defined standardization as a fundamental step towards meaningful transparency. [152] The support for standardization extends beyond the EU, with national and regional regulators advocating for common assessment standards despite the diversity of platforms. [153] This collective endorsement underscores the feasibility and critical need for standardization to ensure that transparency mechanisms serve their intended purpose rather than become mere checkboxes for compliance.

However, the DSA's original provisions did not fully address the

---

[150] *See generally* Halina Szejnwald Brown, Martin de Jong & Teodorina Lessidrenska, *The Rise of the Global Reporting Initiative: A Case of Institutional Entrepreneurship*, 18:2 ENV'T POL. 182, 190-193 (2009), DOI: <10.1080/09644010802682551>; Mikkel Flyverbom, Lars Thøger Christensen & Hans Krause Hansen, *The Transparency–Power Nexus* 29:3 MGMT. COMMC'N Q. 385, 401-402 (2015), https://journals.sagepub.com/doi/10.1177/0893318915593116.

[151] *See, e.g.*, Directive 2013/34/EU of the European Parliament and Council of 26 June 2013 on the Annual Financial Statements, Consolidated Financial Statements and Related Reports of Certain Types of Undertakings, Amending Directive 2006/43/EC of the European Parliament and of the Council and Repealing Council Directives 78/660/EEC and 83/349/EEC, 2013 O.J. (L 182/19) at 3.

[152] European Commission, *Assessment of the Code of Practice on Disinformation - Achievements and Areas for Further Improvement*, (Commission SWD 180) (Sept. 10, 2020) at 21: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=69212. (During the review of the EU Code of Practice on Disinformation, in an attempt to address the inefficacy of the Code).

[153] Zornetta, *supra* note 7.

need for a high level of standardization, pointing to a gap that needs to be bridged. The inconsistency in data presentation across various platforms, as highlighted in Section 3, presents significant hurdles to conducting a comparative analysis and drawing comprehensive conclusions about content moderation practices.[154] This inconsistency is not limited to the metrics used but extends to the formatting of data, further complicating the interpretation and comparison of reports.[155] To move towards actionable transparency, it is imperative that regulators minimize the room for interpretation and guide platforms toward a more unified approach to reporting. This includes standardizing reporting units and formats, clarifying definitions, and providing more detailed breakdowns by content category and Member State, thereby ensuring a baseline level of comparability.

The subsequent sections propose specific amendments to the DSA's transparency mandates in the three focus areas analyzed in this article – Government Orders, Own-Initiative Content Moderation, and Human Resources Involved in Content Moderation – aimed at enhancing their effectiveness and applicability, both within the EU and in jurisdictions contemplating similar regulations.

1. Government Orders

Standardizing how platforms present data concerning government takedown orders and information requests is a step forward in ensuring actionable transparency. Indeed, one aspect that surprised readers of the first DSA reports was the notable discrepancy in the volume of government orders reported before and after the DSA's implementation, with a marked decrease observed post-DSA. This discrepancy is likely not indicative of a reduction in government orders but rather reflects the ambiguity surrounding reporting obligations.

To ensure that transparency mandates meet their goal, ambiguity

---

[154] For example, X reported metrics in both raw numbers and percentage terms (*see* X Report, *supra* note 118, ("TIUC Terms of Service and Rules Visibility Filtering Complaint Overturn Rate"), Facebook only used raw numbers (*see* Facebook Report, *supra* note 119, at 13-16).

[155] Some platforms reported response times in days and others in hours. Similarly, some platforms reported average monthly users as "less than 1M" for smaller user counts, while others offered detailed numbers. *See* Alessia Zornetta, Michael Karanicolas & Nicholas Wilson, *Call for Feedback on the Draft Implementing Regulation Laying Down Templates Concerning the Transparency Reporting Obligations of Providers of Intermediary Services and Providers of Online Platforms* at 3 (Jan. 24, 2024), https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates-/F3451716_en.

should be reduced to a minimum. In this case, the ambiguity stems from the lack of specificity in the provisions, leaving platforms uncertain about when to categorize certain interactions as government orders as opposed to instances that fall under the notice-and-action mechanism. This confusion has inadvertently resulted in a decrease in reported government orders, muddying the waters of transparency and undermining the DSA's objectives. Therefore, further clarity and guidance are needed from the regulators over how government orders are to be categorized and how to distinguish between government entities acting in their governmental capacity or as trusted flaggers. For DSA-like regulatory initiatives, this ambiguity could be avoided by specifically disclosing that government requests submitted through the trusted flagger mechanism should still be categorized under the "government orders" disclosures.

Another problematic aspect of the current reporting lies in its binary approach, which allows platforms to classify orders as either completed or not completed. This oversimplification fails to capture the nuanced spectrum of responses platforms may employ when dealing with government orders. The analysis of post-DSA reports illustrates this shortcoming, highlighting the need for a more granular categorization. To address this issue and enhance the quality of reporting, platforms should be mandated to distinguish among government orders that are rejected, partially complied with, and fully complied with. Implementing such a standardized approach, coupled with more detailed categorization, would significantly improve the clarity and utility of transparency reports, ensuring they accurately reflect the complex interplay between platforms and government requests.

### 2. Own-Initiative Content Moderation

Standardization would also improve reporting on own-initiative content moderation practices, which appears to be the area of greatest interest, especially regarding the use of automated means for content moderation. Once more, the ambiguity created in the DSA provision has led to various interpretations by platforms. More specifically, the lack of definitions for "indicators of accuracy" and "rate of error" has led to a wide range of reporting methodologies, resulting in reports that are difficult to interpret and compare, completely lacking information, and ultimately challenging to meaningfully compare across platforms. Regulators should address this variability to ensure an accurate assessment of the reliability and effectiveness of automated tools.

Providing explicit definitions and guidelines for calculating these indicators will not only standardize reporting practices but also ensure that the reports contribute meaningfully to the overarching goals of the DSA. [156] Foreign regulators considering DSA-like provisions should account for this obstacle when demanding disclosures on indicators of accuracy. A proactive approach would involve convening experts to recommend precise methodologies, which could then be embedded directly within the legal text of accompanying implementation guidelines, thereby circumventing the pitfalls observed in the DSA's implementation.

On a broader scale, it is important to stress that the focus on automated tools in the DSA overshadowed the critical role of human reviewers in the moderation process. Many platforms rely on automated systems to filter content for human reviewers – who ultimately make the final decision – making the accuracy of these human interventions a key factor in the overall effectiveness of content moderation. The DSA transparency mandates completely omit the issue of accuracy for human moderators. To address this gap, regulators should demand that platforms also measure and disclose the accuracy of both hybrid and human-only decisions.

This is even more important for contexts in which content moderation is highly context dependent. Disclosing the accuracy of decisions taken by human reviewers supported by automated tools could hint at gaps in the moderation process and the need to improve the technical resources available to moderators. By incorporating metrics that reflect the performance of human moderators, transparency reports would offer a more comprehensive view of moderation practices, enabling stakeholders to identify areas for enhancement not only in algorithmic accuracy but also in the nuanced judgment calls made by human staff. This approach will provide a more comprehensive view of the moderation process, enhancing the transparency and reliability of the reports.

### 3. Human Resources Involved in Content Moderation

Improving standardization of transparency reports should not be limited at quantitative disclosures. While standardizing qualitative

[156] The predominant recommended methodology appears to be using precision and recall. *See* Johnny Tian-Zheng Wei et al., *Operationalizing Content Moderation "Accuracy" in the Digital Services Act*, Woodstock '18: ACM Symposium on Neural Gaze Detection (Aug. 5, 2024), https://arxiv.org/pdf/2305.09601.pdf (arguing that precision and recall are the best indications of accuracy for the DSA).

disclosures – such as those related to human resources involved in content moderation – is significantly challenging, it is a fundamental step toward ensuring actionable transparency. The varied responses from platforms regarding the training and composition of their moderation teams highlight the need for more precise definitions and consistent reporting standards. This lack of clarity not only impedes meaningful transparency but also hampers the ability to assess the adequacy and effectiveness of the human resources allocated to content moderation.

To improve the quantitative aspects of the reports, regulators should provide clearer definitions – such as clarifying what constitutes a "human resource involved in content moderation" – and leave as little margin for interpretation as possible. The current vagueness has led to disparities in reporting, with some platforms including a wide range of roles in their disclosures – including engineers, policy and legal teams in addition to human reviewers, and others limiting their reporting to content reviewers within the EU.

Furthermore, the varied approaches to reporting linguistic expertise, from proficiency levels to the languages of reviewed content, underscore the need for a standardized framework. Defining the roles to be included in the count of moderation resources and standardizing the reporting according to official EU languages would allow for a more consistent and meaningful comparison across platforms to be achieved, ultimately facilitating a deeper understanding of content moderation practices.

In conclusion, the pursuit of actionable transparency under the DSA necessitates a comprehensive approach to standardization, encompassing not just the metrics and formats of reporting but also the definitions and methodologies employed. By addressing the gaps identified in the current framework and adopting best practices from other domains, the DSA can truly fulfill its potential as a tool for enhancing transparency and accountability in the digital ecosystem.

### B. Auditing Platforms' Disclosures

The need for standardization to ensure actionable and meaningful transparency is matched by the equally critical need for verifying the accuracy and completeness of the information platforms provide. Previous voluntary transparency initiatives, indeed, have been criticized for their lack of verifiability, highlighting a significant gap in accountability. The inability to access a platform's underlying data and

documentation means that stakeholders are often left with an incomplete understanding of content moderation practices. This gap allows platforms considerable leeway to curate their disclosures, potentially omitting or glossing over aspects of their content moderation processes that might cast them in a less favorable light.[157]

The revelations brought to light by Francis Haugen, a former Meta employee, serve as an illustration of this issue. The leaked documents exposed Meta's awareness of its product's detrimental effects on teenagers, the existence of a two-tier justice system providing immunity to "VIP" users,[158] and issues with the accuracy of algorithmic content moderation.[159] Unsurprisingly, such critical issues were absent from Meta's voluntary transparency reports, underscoring the limitations of unverifiable disclosures.

The reliability of transparency reports hinges on the need to grant stakeholders access to the platform's data.[160] This access is pivotal for enabling a straightforward comparison of reported figures against actual data, a process that can uncover discrepancies and enhance the accountability of digital platforms. An illustrative example of the importance of data access can be seen in the case of X, where the platform's DSA transparency report disclosed a total of 19.828 actions taken based on automated content moderation for terms of service and rules violations. At the same time, the statement of reasons submitted to the DSA Transparency Database discloses that not even one decision was made relying on automated means. This discrepancy raises questions about the reliability of both transparency disclosures.[161]

---

[157] *See generally* Monika Zalnieriute, *"Transparency-Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism*, 8 CRITICAL ANALYSIS L. (2021), https://papers.ssrn.com/abstract=3805492 (arguing that the "focus on transparency acts as an obfuscation and redirection from more substantive and fundamental questions about the concentration of power, substantial policies, and actions of technology behemoths").

[158] Jeff Horwitz, *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.*, WALL ST. J. (Sept. 13, 2021), https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353.

[159] Deepa Seetharaman, Jeff Horwitz & Justin Scheck, *Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts.*, WALL ST. J. (Oct. 17, 2021), https://www.wsj.com/articles/facebook-ai-enforce-rules-engineers-doubtful-artificial-intelligence-11634338184 (last visited Feb. 22, 2024).

[160] *Cf.* Christopher Parsons, *The (In)Effectiveness of Voluntarily Produced Transparency Reports*, 58 BUS. & SOC'Y 103 (2019) (on the limitations of companies voluntarily producing transparency reports to promote change in firm and government behavior).

[161] The DSA Transparency Database, launched by the European Commission, is a regulatory platform where online platform providers must publicly share their content moderation decisions. It collects statements of reasons for content removal or access

However, the quest for actionable and meaningful transparency in content moderation extends beyond the mere ability to perform simple data comparisons. While identifying potential mismatches between reported figures and actual instances of content moderation is crucial, it only represents the first layer of a more complex transparency ecosystem.[162] Beyond aggregated data, it is essential to delve into the context within which content moderation decisions are made, including the reasoning behind platforms' policies and the internal guidelines that inform these decisions.[163]

As discussed above, the DSA mandates platforms to provide contextual disclosures, encompassing information about human moderators, internal content moderation procedures, and explanations concerning the purposes and safeguards of automated content moderation tools. Such disclosures are intended to offer stakeholders a more comprehensive view of the platforms' content moderation ecosystem, illuminating the principles and practices that underpin these critical decisions.

Nevertheless, the effectiveness of these contextual decisions hinges on the stakeholder's ability to verify the information provided. Without true access to the platform's operational data and the capacity to independently audit these disclosures, the reliability of the information remains in question.[164] To address this challenge, there is a pressing need for mechanisms that enable independent verification of the platform's transparency reports.

As regulators outside the European Union contemplate adopting regulations akin to the DSA, including its transparency mandates, a critical consideration must be how to ensure the reliability and

---

restrictions as mandated by the Digital Services Act (DSA). *DSA*, *supra* note 2 art 17. *See* Charis Papaevangelou & Fabio Votta, *What Platform Observability Have You Given Us? A First Look into the Statement of Reasons Database* (2024), https://x.com/favstats/status/1760084818099044357 (pre-print forthcoming) ("X was the only platform that always applied a decision on the same day the infringing content was created" and "[s]urprisingly, none of the [statement of reasons] submitted by X specified the use of automated means whatsoever"). *Cf* X Report, *supra* note 118.

[162] Svea Windwehr & Jillian C. York, *Thank You For Your Transparency Report, Here's Everything That's Missing*, ELEC. FRONTIER FOUND. (Oct. 13, 2020), https://www.eff.org/deeplinks/2020/10/thank-you-your-transparency-report-heres-everything-thats-missing; Mark MacCarthy, *Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry* 13–15 (Transatlantic Working Grp. on Content Moderation Online & Freedom of Expression Working Paper, 2020), https://papers.ssrn.com/abstract=3615726 (demonstrating that "[w]hile current platform practices provide real transparency in some regard, the overall insight into platform operations and decision making are limited").

[163] Windwehr & York, *supra* note 163.

[164] MacCarthy, *supra* note 163 at 22–28.

verifiability of transparency disclosures provided by platforms.

The DSA attempts to address this by establishing a data access and scrutiny framework whereby Digital Services Coordinators and the Commission can monitor and assess platforms' compliance with the DSA, including transparency mandates. Additionally, the DSA also places a significant responsibility on researchers and civil society, aiming to facilitate access to data for vetted researchers.[165] Nevertheless, the effectiveness of Article 40 in ensuring the reliability of transparency reports remains to be fully seen.[166] It is crucial that international regulators also identify potential avenues for auditing transparency disclosures.

In designing transparency mandates, it is fundamental to go beyond simply mandating the disclosure of data and to establish clear, enforceable standards for the accuracy and comprehensiveness of the information provided. This may involve setting up auditing bodies and ensuring that researchers have the necessary tools and legal protections to scrutinize and challenge the platforms' disclosures effectively. By addressing these aspects, transparency mandates have the potential to be meaningful and actionable.

## CONCLUSION

The entry into force of the Digital Services Act (DSA) marked a pivotal moment in the evolution of online platform regulation, aiming to usher in a new era of transparency, accountability, and user empowerment. At the heart of the DSA is the commitment to meaningful

---

[165] Article 40 of the DSA imposes obligations on VLOPs and VLOSEs to grant access to data necessary for monitoring compliance with the regulation to competent authorities, specifically the Digital Services Coordinators designated at the national level in the EU Member State of their establishment or the European Commission. This access includes data related to algorithms based on a reasoned request and within a specified reasonable period. Additionally, VLOPs and VLOSEs are required to provide access to vetted researchers for the purpose of conducting research that contributes to the detection, identification, and understanding of systemic risks within the EU, as well as to assess the adequacy, efficiency, and impacts of risk mitigation measures. This obligation entails that platforms may need to explain the design, logic, and testing of their algorithmic systems. *DSA*, art. 40.

[166] Ulrike Klinger & Jakob Ohme, *What the Scientific Community Needs from Data Access under Art. 40 DSA: 20 Points on Infrastructures, Participation, Transparency, and Funding*, WEIZENBAUM INST. (2023); Julian Jaursch, *Here Is Why Digital Services Coordinators Should Establish Strong Research and Data Units*, DSA OBSERVATORY (Mar. 10, 2023), https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/; Pietro Ortolani, *If You Build It, They Will Come: The DSA's "Procedure Before Substance" Approach*, *in* PUTTING THE DSA INTO PRACTICE: ENFORCEMENT, ACCESS TO JUSTICE, AND GLOBAL IMPLICATIONS (2023), https://verfassungsblog.de/dsa-build-it/.

transparency in content moderation practices, a principle that is essential for safeguarding the digital public sphere and reinforcing the democratic values that underpin our societies. Through a comprehensive analysis of the transparency reporting practices of Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) before and after the DSA's implementation, this paper has shed light on the transformative potential of the DSA, as well as the challenges and complexities inherent in achieving actionable transparency.

The analysis revealed that while the DSA has catalyzed advancements in transparency reporting, significant disparities remain in the granularity, consistency, and standardization of disclosures across platforms. These discrepancies underscore the need for a more harmonized approach to transparency reporting, one that enables stakeholders to effectively scrutinize and compare platforms' content moderation practices. Furthermore, the findings highlight the critical role of human resources in content moderation, emphasizing the importance of disclosing not only the scale and composition of moderation teams but also their training, support, and working conditions.

To realize the full potential of the DSA's transparency mandates, this article advocates for the adoption of standardized reporting frameworks akin to those established by the Global Reporting Initiative (GRI) and the International Financial Reporting Standards. Such standardization would facilitate more meaningful comparisons across platforms, enhancing the accountability and efficacy of content moderation practices. Moreover, the article calls for the establishment of mechanisms to verify the accuracy and completeness of platforms' disclosures, ensuring that transparency reports serve as a reliable foundation for regulatory oversight, academic research, and public discourse.

As regulators beyond the European Union look to the DSA as a model for their own digital governance frameworks, the insights gleaned from this analysis aim to offer valuable lessons on the importance of standardization, verifiability, and the comprehensive disclosure of content moderation practices. By embracing these principles, regulators can foster a more transparent, accountable, and democratic digital ecosystem, one that empowers users and upholds the fundamental rights and values that are essential for a thriving digital age.

In conclusion, the DSA represents a significant step forward in the quest for meaningful transparency in the digital realm. However, to fully harness its potential, concerted efforts are required to enhance the

standardization and verifiability of transparency reporting. Through such efforts, the DSA can serve as a beacon for global digital governance, promoting a safer, more accountable, and more democratic online environment for all.