Söderlund, Kasia; Engström, Emma; Haresamudram, Kashyap; Larsson, Stefan; Strimling, Pontus

Article

Regulating high-reach AI: On transparency directions in the Digital Services Act

Internet Policy Review

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

Suggested Citation: Söderlund, Kasia; Engström, Emma; Haresamudram, Kashyap; Larsson, Stefan; Strimling, Pontus (2024) : Regulating high-reach AI: On transparency directions in the Digital Services Act, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 13, Iss. 1, pp. 1-31, https://doi.org/10.14763/2024.1.1746

This Version is available at:
https://hdl.handle.net/10419/296496

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 13 **Issue** 1

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Regulating high-reach AI: On transparency directions in the Digital Services Act

Kasia Söderlund *Lund University* katarzyna.soderlund@lth.lu.se
Emma Engström *Institute for Futures Studies*
Kashyap Haresamudram *Lund University*
Stefan Larsson *Lund University*
Pontus Strimling *Institute for Futures Studies*

DOI: https://doi.org/10.14763/2024.1.1746

Published: 26 March 2024
Received: 21 June 2023 **Accepted:** 7 October 2023

Funding: This work was partially funded by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), Sweden.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

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

Keywords: Digital Services Act (DSA), Social media, AI transparency, Content recommender systems, Artificial intelligence

Abstract: By introducing the concept of high-reach AI, this paper focuses on AI systems whose widespread use may generate significant risks for both individuals and societies. While some of those risks have been recognised under the AI Act, we analyse the rules laid down by the Digital Services Act (DSA) for recommender systems used by dominant social media platforms as a prominent example of high-reach AI. Specifically, we examine transparency provisions aimed at addressing adverse effects of these AI technologies employed by social media very large online platforms (VLOPs). Drawing from AI transparency literature, we analyse DSA transparency measures through the conceptual lens of horizontal and vertical transparency. Our analysis indicates that while the DSA incorporates transparency provisions in both dimensions, the most progressive amendments emerge within the vertical transparency, for instance, by the introduction of the systemic risk assessment mechanism. However, we argue that the true impact of the new transparency provisions extends beyond their mere existence, emphasising the critical role of oversight entities in implementation and application of the DSA. Overall, this study highlights the paramount importance of vertical transparency in providing a comprehensive understanding of the aggregated risks associated with high-reach AI technologies, exemplified by social media recommender systems.

1. Introduction: Transparency of high-reach AI systems

In recent decades, transparency has become a key concept in shaping discussions across a wide range of disciplines, including public governance (Ball, 2009), economic-political debate (Forssbæck & Oxelheim, 2014), as well as digital technology development (Felzmann et al., 2020; Pasquale, 2015) — and its popularity seems to be constantly on the rise (Koivisto, 2022). The prominent role of transparency has also been evident in ethical approaches to artificial intelligence (AI) governance (Jobin et al., 2019), as a crucial element in fostering human-centric and trustworthy AI (High-level Expert Group On Artificial Intelligence, 2019; European Commission, 2020). Increasingly, transparency has been incorporated as a legislative tool in European regulations concerning AI technologies (cf. Larsson, 2021), such as the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and most recently the AI Act (AIA).

In this article, the aim is to deepen the understanding of transparency measures in the DSA specifically, which have been designed to address potential risks arising from what we refer to as *high-reach AI*. As will be elaborated upon below, high-reach AI systems constitute one of the most pervasive and influential technologies in contemporary societies. While such technologies are employed in a variety of contexts, our attention centres on recommendation systems used by dominant social media platforms, which fall within the scope of the DSA.

The scientific literature highlights the potential of recommender systems to inflict various harms, including online nudging that undermines privacy and human autonomy (Yeung, 2017), unintended spread of disinformation (Celliers & Hattingh, 2020), and radicalisation tendencies (Hong & Kim, 2016; Ribeiro et al., 2020). While the extent of the impact of recommendations has also been contested (Barberá, 2020; Jungherr & Schroeder, 2021), there is a consensus in the pressing need for further empirical research to better understand their aggregated effects.

However, the deliberate operational opacity maintained by social media platforms (Rieder & Hofmann, 2020) makes it difficult to evaluate the societal impact of recommender systems. While the activities of social media platforms have been regulated by an umbrella of EU laws, most notably consumer and personal data protection laws, the E-Commerce Directive, and the voluntary Code of Practice on Disinformation (European Commission, 2022a), until recently, their large-scale impact has largely eluded public and regulatory scrutiny. The attempts to conduct independent research concerning the risks associated with social media have been hampered by the strict rules dictated by the platforms' terms and conditions of ser-

vice (e.g. Leerssen, 2021).

As a legislative response to such challenges, the EU regulatory framework for digital services was updated by the Digital Services Act (DSA) (Regulation 2022/2065) in November 2022, replacing the E-Commerce Directive (Directive 2000/31/EC). Among the DSA's many objectives, it aims to mitigate specifically the risks stemming from the large-scale online services by introducing a number of transparency provisions. Although the DSA applies to all intermediary services, the most stringent transparency rules are imposed on very large online platforms (VLOPs) and very large online search engines (VLOSEs), with at least 45 million average EU users per month. The online providers which have been captured by the scope of the social media VLOPs include such platforms as Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, X (formerly Twitter), and YouTube (European Commission, 2023). Although the implementation of the DSA is still pending at the national level, most of the transparency provisions pertinent to VLOPs are already in effect.

The widespread societal adoption of such AI technologies as social media recommender systems can be seen as the defining feature of *high-reach* AI. These broadly used data-driven AI systems have been shown to carry the potential to pose substantial risks for individual users *over time*, as well as *in aggregation* in a way that can exert urgent threats on societies. Examples of high-reach AI include any AI technology used on a large scale, such as music and film suggestions, personalised ads, search engine rankings, and other popular AI technologies like generative AI (ACM, 2023; Lorenz et al., 2023) — that is, AI-models capable of generating texts, images and sounds based on mere prompts by human users.

The emergence of these high-reach AI phenomena clearly links to the digital organisation following from *platformisation* (van Dijck et al., 2018; Poell et al., 2019), often embedded in a particular type of commercial logic (Srnicek, 2017), with not only privacy implications but also antitrust in terms of a few's control over the many (Larsson et al., 2021). In practical terms, high-reach AI is directed towards individuals in their roles as consumers, data subjects, end-users, or recipients of digital services depending on context or regulatory field. We argue that addressing the risks associated with high-reach AI requires the establishment of robust governance frameworks, involving entities granted access to relevant data and equipped with the capacity to assess the large-scale impact of such systems on individuals and societies.

Notably, while the AI Act (European Parliament, 2024) establishes strict gover-

nance rules for AI systems classified as posing high risk to fundamental rights and safety of individuals, the European Parliament's proposal to include social media VLOPs using recommender systems within the scope of high-risk AI systems did not find its way into the final version of the AI Act (European Parliament, 2023, Amendment 740). This category of high-reach AI has, therefore, not been covered by the scope of the AI Act. At the same time, the last changes to the AI Act have introduced a separate compliance regime for the so-called general purpose AI (GPAI) with systemic risks at the Union level, which would encompass the broadly used generative high-reach AI technologies mentioned above (see Recital 99 AIA, European Parliament, 2024). It is worth noting that the AI Act points explicitly to the GPAI's "significant impact on the internal market due to its *reach*" (European Parliament, 2024, Recital 111 AIA; emphasis added). Thus, it appears that the EU lawmakers are increasingly turning their attention to the large-scale impact of the high-reach AI.

Moreover, the rules governing social media recommender systems may be further updated by the EU consumer laws, which are currently under review with regards to such concerns as personalisation practices and addictive use of digital products (European Commission, 2022b).

While acknowledging that social media recommender systems are also subject to other EU legal frameworks, and may fall under the scope of forthcoming regulations, this article focuses on the rules established by the DSA. Our analysis is further limited to the transparency provisions relevant for recommender systems employed by social media VLOPs.

In examining the new transparency rules introduced by the DSA, we draw on the conceptual framework proposed by Heald (2006), which delineates the horizontal and vertical directions of transparency. This theoretical framework serves as our guide in mapping the variety of transparency provisions introduced by the DSA and is helpful in demonstrating that effective legal frameworks governing high-reach AI technologies need to incorporate robust vertical transparency mechanisms.

The article is structured as follows. In the subsequent section, we focus on two types of harms attributed to social media VLOPs as identified in the literature, concerning i) the risk for negative impact on the human end-users' rights to privacy and autonomy, and ii) the risk for amplification of the spread of harmful content online. In Section 3, building on Heald's (2006; 2022) conceptual framework on transparency directions, we establish the theoretical foundation for our analysis

and align the transparency provisions introduced by the DSA accordingly. Moving to Section 4, we discuss the impact of the DSA in mitigating the above risks and what challenges may be expected in this regard. We conclude in Section 5, highlighting the importance of utilisation of the transparency provisions within the vertical direction to foster better understanding of the impact of high-reach recommender systems on both individual and societal levels.

2. The risks posed by social media recommender systems

Recommender systems within social media stand out as one of the most prominent and widespread manifestations of high-reach AI. These systems have become integral to the internet experience (Burke, 2002; Engström & Strimling, 2020; Kozyreva et al., 2021; Xie et al., 2021). All the largest social media services employ recommender systems that are propelled by some form of machine learning technology (i.e., AI) (Engström & Strimling, 2020). They undergird platforms with user bases that often number in the billions, so they are likely to constitute one of the first forms of AI that people encounter in their everyday lives.

In the European regulatory context, the DSA defines a recommender system as:

> a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service, including as a result of a search initiated by the recipient or otherwise determining the relative order or prominence of information displayed. (Regulation 2022/2065, Art. 2)

This definition encompasses a wide range of recommender activities. Such systems tend to go beyond simple ranking by leveraging user data to adapt and personalise suggestions over time (Milano et al., 2020), which can lead to dynamic changes in their function and behavioural impact. Thus, improved understanding of the role that recommender technologies play in today's societies is crucial. Modern democracies rely on well-informed voters with access to relevant and truthful information (Hart et al., 2009).

Recommendations can influence the ability of individuals to access information (Helberger et al., 2021) and as such they may have a large effect on user behaviour. For example, Plummer (2017) reported that 80 percent of what people watch on Netflix is driven by algorithmic recommendations. Given their role in shaping

information flows in the online environment, social media platforms therefore wield significant power as they manage the algorithms that steer recommendation technologies. This motivates well-considered and evidence-based regulations; as the DSA constitutes the most current example of such a policy, we are interested to explore the impacts that this law may have on concerns that have been voiced by the research community.

An important property of personalisation technologies is that they have generally been designed to optimise an aggregate, system-wide measure of performance, such as revenue or accuracy regarding the algorithm's ability to recommend items that are subsequently chosen by the user (Ekstrand & Kluver, 2021). While this can be beneficial for the users in terms of providing more relevant content and mitigating information overload, it can also lead to unintended personal and social outcomes that could emerge over time. There is a risk that individuals are steered towards content that is not beneficial to themselves or to society at large in the long run. Notably, suppliers of recommender technologies have many *different* interests to consider, including short-term obligations to shareholders. In an interview study focusing on news recommendations in China, Xie et al. (2021) found that the decision to use personalisation technology was motivated by increased traffic, engagement, and advertising revenue, which highlights that *other* goals than the users' benefits can underlie the design and adoption of recommender technologies.

This suggests that seemingly insignificant decisions about online content can have substantial *aggregated* impacts, which can be social, environmental, or political in nature. Importantly, such large-scale effects are not observable from the level of individual users, particularly since the systems are often driven by opaque machine learning algorithms. Below, we outline some of them as they have been identified by the research community. While the related empirical evidence is mixed, we argue that the knowledge gap in terms of the social impact of these technologies motivates the need for further research to facilitate evidence-based consensus.

2.1 Impact on privacy and autonomy

The first type of concern we focus on is the negative impact of social media recommender systems on the closely intertwined issues of users' privacy and autonomy. The right to privacy regards the right to keep personal matters secret, free from unwanted access, observation, or intrusion. Altman (1975) defined privacy as "selective control of access to the self or to one's group" (p. 18). Information privacy is

specifically concerned with the control of information about oneself: "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1967, p. 7). Gavison (1980) specified privacy as a concern for limited accessibility, something that regards control over access to the self and protection from unwanted access.

The strong link between privacy and autonomy was further articulated by Zuboff (2019). She argued that external access to personal information undermines the users' autonomy, as the data is used to control their behaviour. She theorised around a system she denoted *surveillance capitalism*, an economic order in which a few internet companies wield extensive powers over online users by managing their data (Zuboff, 2019). This system operates on the premise that access to large amounts of behavioural data coupled with sophisticated inference algorithms empowers corporations to predict and influence user behaviour, which allows for the understanding of how the users' attention can be steered (Zuboff, 2019). For example, this can be based on understanding when, where, and in what form a particular message or advertisement has the largest effect on an individual, or by controlling users in physical space, she argued.

While Zuboff's (2019) theories have been contested (Morozov, 2019), it is unambiguous that social media platforms act in an *attention economy* when they shape the flow of information online (cf. Wu et al., 2017). Also, they clearly have an economic interest to design recommender systems that promote retention and user engagement. This suggests that they have an incentive to collect and store and infer as much personal details as possible. The quality of the collected information improves the more varied, detailed, and recent it is (Friedman et al., 2015). Data can be used to direct the user's attention towards content that is beneficial for the platform, such as advertisements or new features. Also, behavioural data on sites can be cross-linked with other types of information, such as geolocation, purchasing data, browsing history, and search queries, which might then reveal personality features that the user has not (intentionally) consented to sharing (Friedman et al., 2015).

The implication is that the long-term goals of the VLOPs may not always align with their users' interests. It has been argued, for instance, that algorithms which have been designed to increase user engagement can have harmful effects on an individual in the form of internet overuse and social isolation, as well as mental health issues such as depression and anxiety. As reported by Gayle (2021), a Facebook research team found that Instagram aggravates body issues as well as anxi-

ety and depression among adolescent girls. Keles et al. (2020) reviewed evidence on the association between social media use and depression, anxiety, and psychological distress among adolescents (13-18 years). They concluded that there was a general correlation between usage and mental health issues, although the findings were not fully consistent. The most widely reported effect concerned depression. However, Keles et al. (2020) also reasoned that specific behaviours, such as social comparison, might have a larger effect on mental health than usage per se. Although the link between digital technologies and mental health among young people is a debated topic (Bell et al., 2015), other reviews have found a general association between social media use and mental health (Karim et al., 2020). Views among adolescents corroborate this idea (O'Reilly et al., 2018).

While empirical research continues to examine the influence of high-reach technologies online, the negative effects that algorithms have on autonomy has been discussed fervently by normative researchers (Eskens, 2020; Sætra, 2019; Susser et al., 2019; Yeung, 2017). In essence, the right to privacy can be connected to other fundamental rights around the control of information and communication flows, as well as the freedom of expression, opinion and thought. These studies point at the urgency of privacy protection on social media VLOPs.

What emerges from the above empirical and theoretical discourses is that privacy is getting harder to preserve as advanced algorithms are added to recommend features in popular social media platforms (Engström & Strimling, 2020). For the user, there is a compromise to be struck between personalisation and privacy — a phenomenon that has been denoted the *privacy-personalization trade-off* in the literature (Chellappa & Sin, 2005; Friedman et al., 2015). It is clear that the impact of social media recommendations on individuals' rights warrants closer scrutiny and better understanding.

2.2 The dissemination of harmful content

Research on recommender systems raises concerns about their potential to amplify the dissemination of harmful content. Prior studies have demonstrated faster diffusion for affective content (Brady et al., 2017) and false news (Vosoughi et al., 2018), for example. In this context, harmful content is broadly defined, encompassing information that both can cause negative impact *immediately* or over time in the form of aggregate effects. This includes the spread of misinformation (false information regardless of intent), and disinformation (false information with the intent to deceive) (Cambridge Dictionary, 2021), as well as content that may contribute to radicalisation (Ribeiro et al., 2020).

As argued by Celliers & Hattingh (2020), (fake) news creators have an incentive to come up with sensational headlines, which tend to draw a lot of traffic and hence boost ad revenue, because the main business model on the internet is advertising. Correspondingly, Vosoughi et al. (2018) showed that false news items tend to spread more easily than true ones on Twitter (now X); the former could reach audiences that were more than 100 times larger than the latter, which illustrates the high-reach aspect of social media recommender features. There are also concerns that generative AI is adding to this spread of mis- and disinformation (Lorenz et al., 2023).

Bak-Coleman et al. (2021) reasoned that our limited understanding of how social media changes our social systems challenges democracy, scientific progress, and our ability to address urgent global problems such as pandemics. Lührmann et al. (2020) identified three threats linked to the proliferation of illiberalism and authoritarianism online: the spread of disinformation, filter bubbles, and hate speech. Filter bubbles refers to the notion that online users by algorithmic mediation tend to be exposed to views and information that confirm their prior understanding (Pariser, 2012). In early work on this topic, Van Alstyne & Brynjolfsson (1996) warned that information technologies could lead to *cyberbalkanisation*, the gathering of individuals in clusters of politically like-minded others.

It has been argued that personalised information consumption online has been a key contributor to the fragmented political discourse and increased social polarisation that we have seen in recent years. Facebook's recommender engine spurred some users towards extreme media content, as shown in the company's internal reports (Zadrozny, 2021). The platform's recommender algorithms had provoked 64% of extremist group joins (Horwitz & Seetharaman, 2020). Ribeiro et al. (2020) analysed YouTube data and found that users migrated towards more extreme channels over time. Similarly, a former YouTube engineer has claimed that the platform's recommendation algorithm promotes conspiracy theories (Turton, 2018). This has been corroborated by other studies; for example, in a meta-analysis on the YouTube recommender system, Yesilada and Lewandowsky (2022) reported that 14 out of 23 studies found empirical evidence supporting the notion that the system could lead users towards problematic content (seven found mixed results, and two did not find evidence of such content pathways).

In a literature review, Barberá (2020) examined the association between social media and political polarisation, identifying emerging consensus and open questions. Barberá reported that evidence suggested that social media usage increased the diversity of views that individuals were exposed to, although findings around the

implications of this were mixed. For example, Boxell et al. (2017) found that the increase in political polarisation in the US was the lowest among those who were most likely to use social media, which suggested that usage did not have a substantial effect on polarisation. On the other hand, Bail and colleagues (2018) reported that individuals who were exposed to views that opposed their own on Twitter became more polarised; specifically, Republicans became more conservative after receiving updates from a liberal Twitter bot. A study by Allcott et al. (2020) found that Facebook deactivation decreased polarisation of views around policy issues, implying that usage had a dividing effect.

In conclusion, while the societal effects of social media recommendation systems have been disputed, we argue that there is ample reason to be concerned about the algorithms that drive such technologies. This motivates the need for much more independent and evidence-based research on high-reach AI. Notably, Yesilada and Lewandowsky (2022) highlighted that their analysis was incomplete due to their limited access to the YouTube recommender system. Barberá (2020) and Leerssen (2021) problematised that researchers have been unable to access ranking algorithms in social media, and that the most extensive studies in the field tend to be conducted by the platforms themselves. These are clearly not independent and this brings their trustworthiness into question (Urman & Makhortykh, 2023).

Many of the transparency provisions introduced by the DSA are intended to address these challenges. Below, we present the analysis of the DSA provisions by using Heald's (2006) *analytics of transparency* as a guiding framework to explore how the different directions of information flow can contribute to addressing these issues.

3. The transparency directions in the DSA

Despite the extensive discussions on transparency across various domains (Koivisto, 2022), it remains a vaguely defined and multifaceted concept (Larsson & Heintz, 2020). As Alloa (2022) observes, the ubiquity of transparency rhetoric in contemporary discussions spanning media, politics, industry, finance, and technology markedly contrasts with its inherent lack of systemic clarity. Meijer (2014) claims that transparency is an 'ideograph', a concept that is seemingly universally desirable yet inherently hollow, allowing it to be framed in various ways depending on the context.

In light of this conceptual ambiguity, various nuanced approaches to transparency

discussions have been proposed, emphasising the need for more effective and contextually appropriate transparency strategies. For instance, Meijer (2014) identifies 'three basic perspectives' on transparency: as (1) a virtue, (2) a relation, and (3) a system. When defined as a virtue, transparency signifies being open about one's behaviour, intentions, and considerations, without specifying the audience to whom this transparency is directed. When defined as an institutional relation, transparency is an exchange of information between 'an actor and a forum'. When understood as a system, transparency is a network of relations, each transparent on their own level, contributing to the transparency of the whole (Meijer, 2014).

Another insightful perspective on transparency has been proposed by Heald (2006). In his examination of *anatomy of transparency*, Heald emphasises the need to consider factors such as the direction of information flow, different varieties of transparency (event transparency versus process transparency, transparency in retrospect versus transparency in real time, etc.), and the specific habitats (contexts) in which transparency operates.

In this study, we focus on Heald's conceptualisation of transparency directions, which offers a helpful theoretical scaffold in dissecting the transparency measures provided by the DSA. The framework distinguishes between (1) upward, and (2) downward transparency — which together form vertical transparency, (3) inward, and (4) outward direction — which together form horizontal transparency. In the horizontal direction, inward transparency allows those outside a system to *view in,* and outward transparency allows those inside the system to view their outside environment. In the vertical direction, upward transparency pertains to transparency directed towards a hierarchically superior body, and downward transparency allows the hierarchically inferior to monitor the governor by virtue of their position as the observed or governed. When all transparency directions are equally present, it is referred to as *fully symmetric transparency*, whereas the absence of all transparency directions results in *fully symmetric non-transparency*. Moreover, some transparency directions may be asymmetrical or completely absent (Heald, 2006). In the upcoming subsections we employ this framework by situating it within the context of digital services and allocate the transparency measures provided by the DSA accordingly.

3.1 Horizontal transparency

In line with Heald's approach, the interaction between actors in the horizontal transparency directions occurs without the involvement of a hierarchically superior (regulatory) body. Figure 1 illustrates the model of horizontal transparency be-

tween the platforms and end-users:

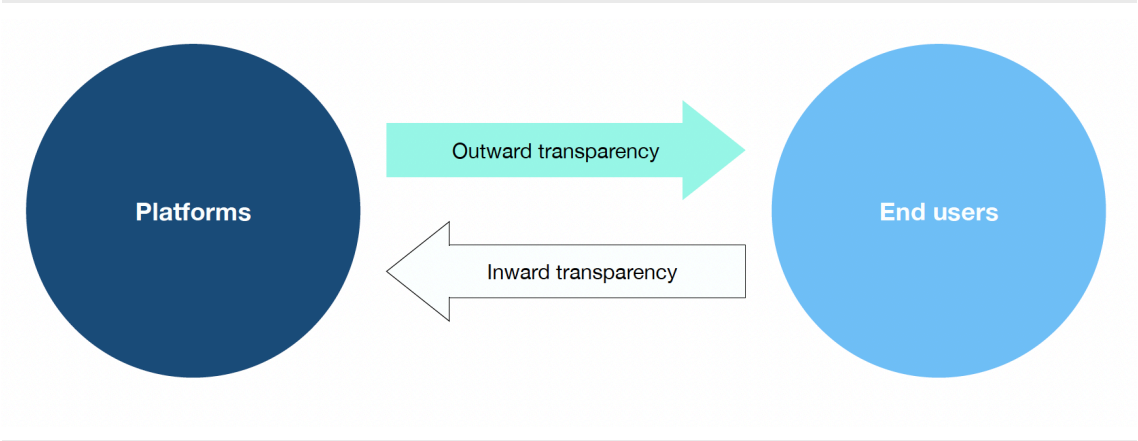

FIGURE 1: The model of horizontal transparency.

3.1.1. Inward transparency

In general, inward transparency can be understood as the capacity to be observed or monitored by entities considered as equal from a legal standpoint. This characterises the relationship between platforms and end-users, established in most cases through a contractual agreement. Inward transparency can thus be seen as available means for users or the general public to examine the platforms' internal workings and practices.

In the DSA, inward transparency measures aimed at allowing end-users to *view in* may include general information that providers publicly disclose about their services, as well as information that must be provided to end-users in a specified manner mandated by relevant laws. The providers typically fulfil this obligation through their terms and conditions of service and privacy policies. Although the rules concerning provision of information about the digital services can be found in EU consumer and data protection laws, and certain specific information was also required by the E-Commerce Directive, the DSA further expands the information provision list. For instance, users must now be informed about content moderation practices and usage restrictions, including the prohibition of certain types of content (Regulation 2022/2065, Art. 14).

Another novel inward transparency provision states that VLOPs are obliged to disclose the *main parameters* used in their recommender systems and explain why the specific information is suggested to the user. Following Article 27, this includes the "criteria which are most significant" in determining the presented content, and "reasons for the relative importance of those parameters" (Regulation 2022/2065).

Additionally, the DSA creates new avenues for user-provider interactions, such as in the involvement of end-users in the mechanism of tackling illegal content. Users may flag what they consider as illegal in the so-called *notice-and-action* mechanism. The platforms are obliged to consider each notification, and their decisions in this respect may range from the restrictions in visibility of content by its removal, disabling access or demotion, to suspension or termination of the user's account. Importantly, following Article 17 of the DSA, the users affected by the platform's action should be provided with a *statement of reasons* (grounds for such decisions). Moreover, the online platform should provide the users, both the ones notifying the illegal content and the ones negatively affected by the platform's decision, with the possibility of access to an internal complaint handling and out-of-court dispute settlement (Regulation 2022/2065, Art. 20 and 21 respectively).

3.1.2. Outward transparency

On the flip side, outward transparency in the digital services domain can be interpreted as the provider's activity of observing any relevant phenomena in its environment, including the (planned) actions of the regulators, competitors, and market trends. Crucially, it also involves closely monitoring the users of the digital services.

The role of regulations in outward transparency is mostly focused on curbing the practice of monitoring end-users by extensive collection of personal data, which, as pointed out earlier, may amount to the intrusion on users' privacy. While the GDPR and consumer law already forbid certain practices, the DSA adds more targeted provisions in this regard, such as by obliging the VLOPs to enable an option to use recommender systems without being profiled (Regulation 2022/2065, Art. 38).

Moreover, as regards advertising, it seems that the DSA reiterates an already established rule under the GDPR, by prohibiting advertising based on the use of sensitive data in the meaning of the GDPR, such as revealing racial or ethnic origin, political opinions, data concerning health or sexual orientation (Regulation 2022/2065, Art. 26(3)).

Nevertheless, as elaborated upon in Section 2, the extent of the engagement in outward transparency activities by high-reach platforms raises significant concerns. As will be further discussed in Section 4, while end-users have certain transparency tools at their disposal, these fall significantly short of the level of monitoring which end-users are subjected to themselves. Using Heald's terminology, the horizontal transparency is thus highly *asymmetric* and imbalanced, which necessi-

tates the regulatory intervention in form of the vertical dimension of transparency.

3.2 Vertical transparency

In turn, following Heald's framework, the vertical transparency model involves hierarchical relationships commonly used in economic modelling or public governance (Heald, 2006). In the context of digital services governance this relationship can be illustrated as follows (Figure 2):

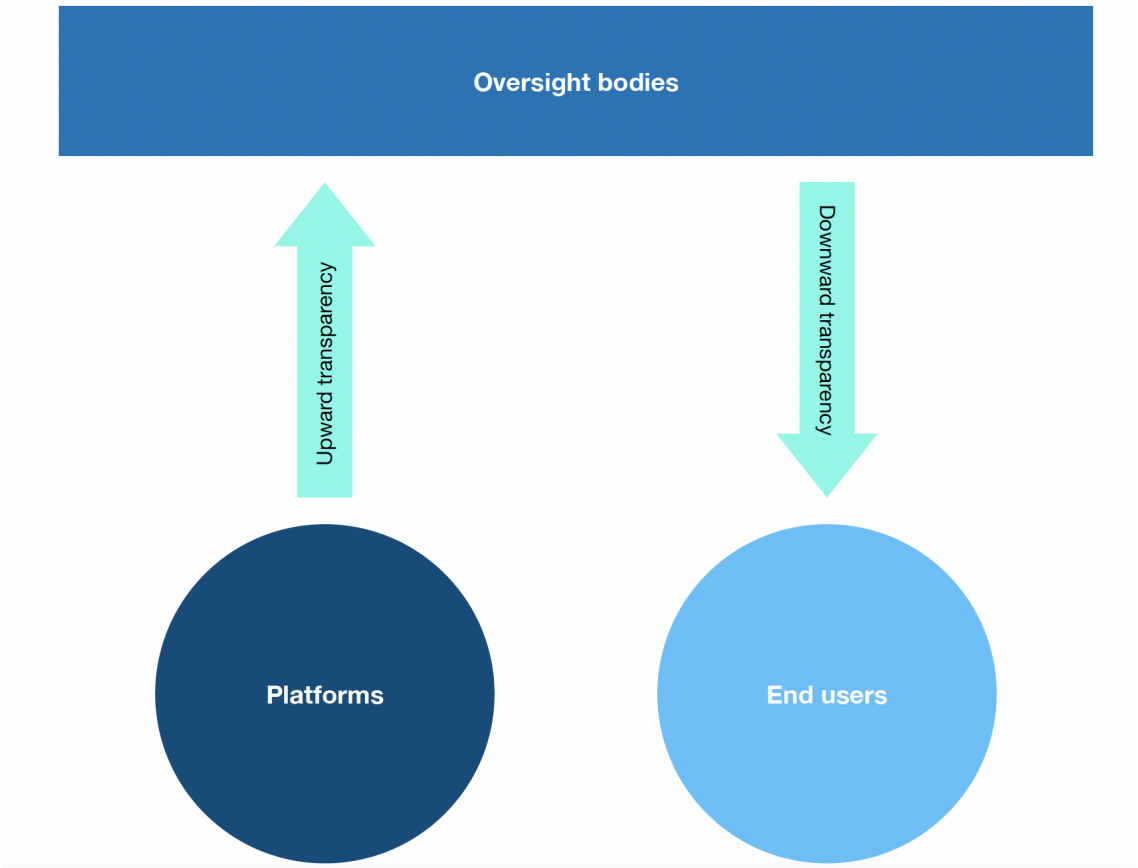

FIGURE 2: The model of vertical transparency.

Upward transparency direction refers to the observability of a subordinate to a hierarchical superior Heald (2006). In state governance, upward transparency is present to varying degrees in all functioning states (Heald, 2022). In the realm of digital services, this activity could be exercised by public authorities with vested responsibility to monitor the activities of online providers.

3.2.1 Upward transparency

Prior to the adoption of the DSA, oversight activities concerning high-reach social media platforms were primarily conducted by national competent bodies within

the regulatory frameworks encompassing digital services, data protection, and consumer protection laws.

However, the DSA has shifted the oversight responsibility over VLOPs to the European Commission, entrusting it with the authority to act in cases of suspected breaches of the DSA (Regulation 2022/2065, Art. 65(2)). On the national level, the Digital Service Coordinators (further referred to as DSCs), whose establishment is currently underway, will be the national authorities tasked with overseeing and enforcing the DSA domestically (Regulation 2022/2065, Art. 49(1)). Moreover, the DSA establishes the European Board for Digital Services, an independent body composed of Digital Service Coordinators which primarily serves in a supportive, coordinating, and advisory capacity (Regulation 2022/2065, Art. 61-63).

Crucially, in the context of upward transparency, the DSA introduces several new upward transparency tools for the above oversight authorities and the most noteworthy provisions will be outlined below.

Assessment of systemic risks

One of such important upward transparency measures requires the VLOPs to conduct what is termed *systemic risk assessments*. The outcomes of such assessments should be provided "upwards" to the Commission and the respective DSCs. The systemic risk assessment mechanism can be outlined as follows:

1. At least once per year, the VLOPs are obliged to identify "any significant systemic risks stemming from their algorithms", such as dissemination of illegal content, any risks for the EU fundamental rights (including risks to privacy, non-discrimination, consumer protection), any serious effects on users' physical or mental well-being, as well as their impact on civil discourse, electoral processes, public security and public health (Regulation 2022/2065, Art. 34).
2. Next, the VLOPs are required to implement "reasonable, proportionate and effective mitigation measures" corresponding to the identified risks, for example by adapting the design of their services, improving content moderation processes, testing and adapting their algorithmic systems (Regulation 2022/2065, Art. 35).
3. Both the assessment and mitigation measures are then subject to *independent auditing*, with the VLOPs providing necessary access, cooperation, and assistance to auditors (Regulation 2022/2065, Art. 37).
4. An audit report is produced, along with the auditor's opinion as to whether the audit is 'positive,' 'positive with comments,' or 'negative'. In the latter two cases, the auditor provides *operational recommendations* that the VLOP must address within a specified period. The VLOP must then submit an

audit implementation report within one month.
5. Finally, the VLOPs publish the transparency report containing results of the systemic risk assessment, implemented mitigation measures, audit report, and audit implementation report. In case certain confidential or sensitive information has not been included in the publicly available version, the full report should be shared with the Commission and the DSCs.

However, as will be shown in Section 4, it should be noted that the effectiveness of this transparency mechanism largely depends on the way the Commission implements it through delegated acts, specifying the rules for the performance of the audits (Regulation 2022/2065, Art. 37(7)).

Access to data

Another noteworthy upward transparency measure in the DSA is that the DSCs or the Commission may request VLOPs to provide them with access to data that are necessary to monitor and assess compliance with the DSA (Regulation 2022/2065, Article 40). The providers may also be requested by the same institutions to explain the design, logic of the functioning and the testing of their algorithmic systems, including their recommender systems.

Importantly, the DSCs may also enable the data access for *vetted researchers* (recognised on the basis of the criteria specified in the DSA) to conduct the research that "contributes to the detection, identification and understanding of systemic risks in the Union, … and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures" which have been applied by the provider (Regulation 2022/2065, Article 40). This is arguably the first legislative initiative providing the research community with access to platform data by their involvement within the DSA governance framework (Leerssen, 2021). It could, notably, be seen as serving the dual purpose of inward and upward transparency, since on the basis of this provision such independent research would benefit not only the general public but also the auditors (Regulation 2022/2065, Rec. 97) and the oversight authorities (Regulation 2022/2065, Rec. 137).

Nevertheless, this measure is subject to significant caveats. First, the VLOPs may refuse to provide data access to vetted researchers under the claim that they do not have access to such data themselves, or that giving access might lead to "significant vulnerabilities for the protection of confidential information, in particular trade secrets" (Regulation 2022/2065, Art. 40 (5)). In such cases, the platforms are required to ask the DSCs to amend the request, and propose an alternative means, which would be appropriate and "sufficient for the purpose of the request". Second,

the Commission is expected to adopt delegated acts supplementing these rules, by specifying the procedural and technical conditions under which the VLOPs are to share the data in compliance with the GDPR, including the purposes for which the data may be used (Regulation 2022/2065, Art. 40 (13)). Thus, as in the case of systemic risk assessments, the effectiveness of this provision is dependent on its interpretation by the platforms, the DSCs, and the Commission.

Investigation powers of the Commission

As the utmost form of upward transparency, the DSA has granted the Commission extensive investigation powers in relation to VLOPs, as outlined in Chapter IV of the DSA. This authority enables the Commission to perform various actions when necessary, such as the power to:

1. Require information from VLOPs, auditors, and individuals or entities who may possess relevant information regarding suspected infringements.
2. Initiate proceedings in case of a suspicion of violation of the DSA.
3. Conduct unannounced inspections of any premises of the above entities or persons, allowing them to examine records and books, make copies, and seal any premises.
4. Require explanations regarding the organisation, functioning, IT system, algorithms, data-handling, and business practices.

The investigations conducted by the Commission can ultimately lead to enforcement actions, ranging from ordering interim measures, imposing fines or periodic penalty payments, to initiating the procedure of restricting access to the service, when all other efforts to stop the infringement of the DSA will have been exhausted.

3.2.2 Downward transparency

Lastly, the downward transparency direction characterises the relationships between the state and various actors (especially citizens) within democratic societies (Heald, 2006). It is often associated with accountability and is a prominent feature of democratic theory and practice, distinguishing it from totalitarian regimes (Heald, 2022). This transparency direction entails that end-users and the general public may ”watch the watchers” by holding regulatory authorities accountable in their responsibility of oversight.

The DSA provides several tools which can be interpreted as downward transparency measures. For instance, the European Board for Digital Services, in cooperation with the Commission, has the responsibility to produce comprehensive yearly re-

ports compiling transparency reports submitted by the digital service providers, including the systemic risk assessments. These reports may also draw from other sources, such as the analyses of the data accessed by the DSCs or vetted researchers (Regulation 2022/2065, Art. 35). Moreover, when the Commission exercises its investigative powers, decisions concerning non-compliance of providers, interim measures, or imposed fines will also be publicly available (Regulation 2022/2065, Art. 80). Furthermore, certain downward transparency measures are also relevant for vetted researchers, who by their involvement in the DSA framework bear the responsibility towards the public for their work. In this regard, the research results are expected to be publicly available and free of charge (Regulation 2022/2065, Art. 40).

In sum, a well-balanced and symmetric vertical transparency entails that governing bodies may hold the providers-platforms accountable for their actions. On the other hand, citizens, as end-users, may hold the governing bodies accountable, primarily the DSCs and the Commission, for their oversight tasks. This reciprocal accountability mechanism aims to establish a system of checks and balances in the online platform governance (Djeffal et al., 2021), preventing abuses of power and ensuring that both platforms and oversight bodies are held responsible for their actions.

4. Addressing the risks posed by high-reach AI through transparency directions

Applying Heald's *anatomy of transparency* to the context of digital services showcases the highly asymmetric and imbalanced information flows across the horizontal transparency directions. Outward transparency, manifested through the extensive engagement of social media platforms in collecting and analysing personal data on a massive scale raises significant concerns as regards the users' privacy, autonomy, and the proliferation of harmful content across societies. In contrast, the inward transparency measures available to end-users, allowing them to understand and exert control over social media recommender systems, are considerably limited.

However, in what appears to be a shared characteristic of high-reach AI systems, relying on the horizontal transparency measures alone is insufficient to rebalance this information asymmetry. To address the large-scale risks posed by social media recommender systems it is necessary to establish the vertical transparency governance dimension for several reasons. Firstly, the technical complexity and legal protection surrounding the high-reach recommender systems make it difficult for

average end-users to determine their trustworthiness (Burrell, 2016). Secondly, even if end-users had more transparency tools at their disposal they would not have the ability to check that the information provided is true, considering the past instances where platforms such as Facebook have made false claims about privacy policies (Pasquale, 2015). Thirdly, given the users' disconnection from power, the corrupt practices may continue even if made known to the users (Ananny & Crawford, 2018). Fourthly, as argued in this paper, some risks may only become apparent on an aggregated level, impacting groups or society as a whole, rather than individual end-users (cf. Knowles & Richards, 2021). Thus, the reliance on the horizontal transparency alone might lead to *transparency fallacy* or *transparency illusion* (Edwards & Veale, 2017; Heald, 2006). Consequently, a stronger emphasis on vertical transparency directions appears necessary to calibrate the asymmetry of information power in the digital services domain.

The adoption of the DSA holds the potential for certain improvements in this regard. The analysis of the provisions introduced by the DSA reveals that the transparency measures aimed at addressing concerns related to privacy, autonomy, and the dissemination of harmful content are present in all transparency directions. However, the extent to which these requirements are incorporated varies across the different transparency directions and arguably may differ in their expected impact on the overall governance of the social media recommender systems.

4.1 DSA and the right to privacy and autonomy

As has been shown, the issue of privacy protection highlighted in Section 2.1 directly conflicts with business models of VLOPs that are based on the collection and analysis of extensive user data, allowing the platforms to create detailed user profiles and increasingly accurate recommendations. The related issue of limited user autonomy is also concerning, as empirical studies have indicated that accurate user profiles can make individuals susceptible to nudging and manipulation. It seems that the risks associated with recommender systems go beyond specific groups. As Helberger et al. (2022) observe, the exerted influence of high-reach social media can make users, to varying degrees, vulnerable to such systems.

Following Heald's transparency framework, the DSA provision requiring the disclosure of the *main parameters* used in VLOP's recommender systems can be interpreted as an inward transparency tool. However, its practical impact remains to be seen, considering that most users do not read terms and conditions of services (Larsson et al., 2021). In addition, the definition of the "main parameters" is unclear. For this reason, the European Data Protection Supervisor (EDPS) called for

specifying what parameters would need to be disclosed at a minimum to constitute "meaningful information" in this context (EDPS, para 76).

To some extent, the notice-and-action mechanism will enhance inward transparency by providing users with the specific grounds for the platform's content moderation decisions. However, in order to avoid accusations of censorship or bias, platforms often resort to more subtle demotion techniques (Gillespie, 2022). As the notice-and-action procedure will generate an increased workload for platform moderation (e.g. due to the internal complaint handling), this may potentially lead the platforms to use demotion techniques that are more difficult to detect, thus providing an incentive to foster an environment where content moderation plays out through dynamic ranking systems (cf. Leerssen, 2023).

Moreover, the DSA stipulates that end-users *may* be given the possibility to modify the main parameters apart from the mandatory, non-profiling option. Although this could arguably enhance the level of user autonomy it is currently not compulsory for the platforms. While many users would appreciate the possibility to customise their algorithmic preferences, it is yet to be seen whether the dominant platforms will make such options available and how this would work in practice. Indeed, Helberger et al. (2021) observe that the VLOPs currently do not have incentives to provide such alternative options for users. It is noteworthy that the proposed interoperability options, which would facilitate the use of recommender systems prioritising individual preferences, were not included in the final text of the DSA. Also in *this* context, the EDPS advocated for introducing minimum interoperability requirements for VLOPs, as this would increase the potential for the development of a "more open, pluralistic environment, as well as create new opportunities for the development of innovative digital services" (EDPS, 2021, para. 84). As Helberger and colleagues (2021) argue, such provision would constitute a true step towards curbing the dominant power of "Very Large Online Goliaths" in the digital services domain.

In turn, within the outward transparency direction, the DSA introduces two noteworthy privacy-enhancing amendments. First, it makes it mandatory for VLOPs to create a usage option which is not based on collecting user data, and notably, as emphasised by EDPS (2021), this should be a default setting following the GDPR *privacy by default* principle (Regulation 2016/679, Art. 25(2)). Arguably, the purpose of this provision is to clarify that the creation of detailed user profiles is not *essential* for providing the service itself (Regulation 2016/679, Art. 25(2)). Although online services primarily generate revenue through advertising, there are potentially less privacy-intrusive alternatives such as context-based advertising. However, the

extent to which users will choose this alternative remains uncertain given the "sticky design" of algorithmic feeds (Haugen & Committee on Internal Market and Consumer Protection, 2022) and the fact that privacy-conscious individuals can already utilise many platform services without logging in. Furthermore, the DSA seems to reiterate the prohibition of advertising based on the use of sensitive personal data, as defined by the GDPR (cf. Regulation 2022/2065, Art. 26(3) and Regulation 2016/679, Art. 9), since advertising does not fulfil any of the legal bases allowing the processing of such sensitive data categories. Nevertheless, both these clarifications have presumably been added due to the apparent proliferation of such practices.

Finally, the DSA transparency provisions aiming to enhance the users' protection of right to privacy and autonomy have been identified within the upwards transparency direction as well. Most notably, the negative impact of social media recommender systems on the EU fundamental rights, such as risks the right to privacy, non-discrimination, and consumer protection are included in the systemic risk assessment reports. Moreover, on the basis of Article 40 of the DSA, the research community may also have the possibility to conduct research in this respect. As has been signalled above however, the actual impact of these transparency mechanisms is dependent on their interpretation and implementation by the oversight bodies.

4.2 DSA and the dissemination of harmful content

As presented in Section 2.2, high-reach recommender systems that have been designed to optimise user engagement may contribute to the amplification of the spread of harmful content online including false news and conspiracy theories. Moreover, personalised information consumption that has been facilitated by such algorithms may add to the fragmentation of the political discourse in democracies. Thus, due to the limited ability of individuals to shape the large-scale effects of social media, the measures aimed at mitigating such aggregated risks can only be effectively addressed by establishing a governance framework with entities capable of exercising effective oversight.

Indeed, the analysis of the DSA provisions concerning the negative societal effects of social media platforms suggests that the most significant changes have been introduced within the upward transparency direction. The systemic risks assessment, including mitigation measures, independent audits and reporting, have the potential to offer valuable insights to both the Commission, DSCs, and the general public. As Rieder and Hofmann (2020) observe, holding platforms to account requires

'institution-building', which necessitates the development of skills and competence "in a form that transposes local experiments into more robust practices able to guarantee continuity and accumulation" (p. 23). Recognising this need, the Commission has established the European Centre for Algorithmic Transparency (ECAT) (European Commission, n.d.), which is tasked to collaborate with industry representatives, academia, and civil society organisations to, among other objectives, evaluate risks and recommend transparency approaches.

However, only the future holds the answer to the question of the actual effectiveness of the risk assessment mechanism as it largely depends on the way these rules will be implemented in practice. The Commission's delegated acts seem essential in this regard as they will specify the rules for the performance of the audits (Regulation 2022/2065, Art. 37(7)). Since transparency reports produced by platforms do not provide meaningful transparency (Urman & Makhortykh, 2023), the establishment of detailed procedures for audits is critical for the efficacy of the framework. Moreover, the need for detailed industry standards serving as benchmarks against which to audit the algorithms have been voiced by some of the auditing corporations themselves. Without such standards, the auditors would essentially find themselves in the position of "DSA judges" determining what qualifies as DSA compliance (Bertuzzi, 2023).

Likewise, the inclusion of vetted researchers within the institutional framework potentially creates the possibility for the research community to actively participate in the monitoring tasks of VLOPs by offering independent expertise to the governing bodies, as well as the general public. While this provision, theoretically, holds great potential, it also contains significant limitations, as mentioned in Section 3.2. VLOPs may refuse data access requests by claiming they do not possess the data, or where access would pose "significant vulnerabilities" to security or "protection of confidential information, in particular trade secrets". As Leerssen (2021) notes, platforms such as Facebook have already abused privacy regulations as a justification for denying researchers access to data, thus similar strategies are likely to be applied with regards to the use of security and trade secrets considerations. Moreover, a practical constraint of this provision could also be the need for vetted researchers to have sufficient technical resources and infrastructure at their disposal to analyse the vast volumes of data involved (Rieder & Hofmann, 2020). Furthermore, just as seen in the instance of the systemic risk assessment mechanism, the success or failure of this ambitious transparency provision hinges upon the Commission's delegated acts in specifying the rules and procedures for the data sharing.

Finally, the Commission's extensive investigation powers serve as a “last resort” measure of upward transparency. Drawing a parallel to EU competition law, state aid, and environmental law, the Commission may initiate proceedings against VLOPs, request information, conduct inspections, and ultimately enforce measures such as imposing fines and/or periodic penalty payments. It could be argued that the direct oversight of the Commission over the VLOPs could be a lesson learned from the shortcomings within the data protection enforcement, where a few authorities are responsible for the majority of EU supervision (Holznagel, 2021). Nevertheless, the effectiveness of the Commission's broad array of investigative powers primarily depends on its proactive use of these enforcement tools.

In sum, while prior to the introduction of the DSA the *transparency asymmetry* was profound, the question remains to what degree this regulatory intervention can recalibrate the information and power imbalance between end-users and VLOPs.

As has been argued, the horizontal transparency measures alone are not likely to effectively address the aggregated impact of high-reach AI, which has been demonstrated in the example of the social media recommender systems. This capability could be vested in entities with access to comprehensive datasets, possessing the adequate expertise and enforcement tools. Although vetted researchers are now involved in a more formalised manner within the DSA framework, effective oversight of the high-reach recommender systems would require regulatory powers substantial enough to shape the actions of the dominant VLOPs. Nevertheless, the caveats surrounding the upward transparency measures within the DSA indicate that the actual impact of the DSA remains uncertain as it largely depends on the interplay of the power dynamics between the regulators and the platforms.

5. Conclusions

High-reach AI systems are increasingly becoming an integral part of modern societies, presenting challenges on both individual and societal levels. As AI technologies are still rapidly evolving, new high-reach AI applications are expected to emerge.

While recommender systems used by social media, as one of the examples of such high-reach AI technologies, offer various benefits for users, they have also been shown to come with inherent risks. These algorithmic systems are primarily optimised for the advertising-driven business models of online platforms, and the lack of effective regulation and enforcement allowed the platforms to flourish without

adequately addressing the risks and harms that they generate. The adoption of the DSA represents a significant and long-awaited step towards addressing these issues by introducing new transparency tools and providing the possibility for more comprehensive oversight and, when necessary, enforcement.

The analysis in this article has focused on the transparency provisions provided by the DSA and has examined them through the lens of the horizontal and vertical transparency framework (Heald, 2006). Our findings indicate that along the horizontal transparency axis end-users can expect to be provided with additional information, such as on content moderation and the main parameters used in recommendations. The most notable amendment in this respect appears to be in the formalising of the *notice-and-action* procedure, which requires the provision of explicit grounds of the decisions that directly affect individual users. However, the proposal for the introduction of interoperability measures, which would provide the users with meaningful choice in their recommendation preferences, have not been included in the scope of the DSA.

Within the vertical dimension, the DSA holds a potential to strengthen the upward transparency, especially with regards to addressing the risks of large-scale harms caused by the social media recommender systems. The EU Commission has been granted extensive investigation powers in this regard, similar to its role in the EU competition, state aid, and environmental law. Nevertheless, among the most innovative transparency tools in addressing the risks posed by social media VLOPs are the provisions concerning the systemic risk assessments and data access for researchers. Yet the latter is subject to significant limitations, primarily grounded in issues of confidentiality and trade secrecy, which may potentially undermine the main objectives of this transparency mechanism. Crucially, both mechanisms, the risk assessment and the data access for vetted researchers, are highly dependent on their implementation through the Commission's delegated acts which are central to the effectiveness of these provisions.

Addressing the risks posed by high-reach social media by the introduction of the DSA serves as an example of the challenges inherent in regulating high-reach AI systems. As has been argued, the monitoring of the broad-scale effects of such AI systems necessitates the establishment of effective governance frameworks, in which vertical transparency plays a pivotal role, offering the means for a holistic overview and regulatory tools to intervene in cases it is needed. Without the operative vertical transparency dimension the aggregate impacts of high-reach AI may go unnoticed. While the DSA holds great promise in recalibrating of the transparency asymmetry in the digital services domain, the actual impact of this new

regulatory framework largely depends on the way the rules are implemented and utilised by the oversight bodies.

References

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–676. https://doi.org/10.1257/aer.20190658

Alloa, E. (Ed.). (2022). *This obscure thing called transparency: Politics and aesthetics of a contemporary metaphor*. Leuven University Press. https://doi.org/10.2307/j.ctv26dhjc9

Altman, I. (1975). *The environment and social behavior: Privacy, personal space, territory, crowding* (1st ed.). Brooks/Cole Publishing Company. https://books.google.com/books/about/The_Environment_and_Social_Behavior.html?hl=sv&id=GLBPAAAAMAAJ

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

Association for Computing Machinery. (2023). *Principles for the development, deployment, and use of generative AI technologies* [Statement]. ACM Technology Policy Council. https://www.acm.org/articles/bulletins/2023/july/tpc-principles-generative-ai

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., Donges, J. F., Galesic, M., Gersick, A. S., Jacquet, J., Kao, A. B., Moran, R. E., Romanczuk, P., Rubenstein, D. I., Tombak, K. J., Van Bavel, J. J., & Weber, E. U. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), Article e2025764118. https://doi.org/10.1073/pnas.2025764118

Ball, C. (2009). What is transparency? *Public Integrity*, *11*(4), 293–308. https://doi.org/10.2753/PIN1099-9922110400

Barberá, P. (2020). Social media, echo chambers, and political polarization. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy* (pp. 34–55). Cambridge University Press.

Bell, V., Bishop, D. V. M., & Przybylski, A. K. (2015). The debate over digital technology and young people. *BMJ*, Article h3064. https://doi.org/10.1136/bmj.h3064

Bertuzzi, L. (2023, July 24). Europe enters patchy road to audit online platforms' algorithms. *Euractiv*. https://www.euractiv.com/section/platforms/news/europe-enters-patchy-road-to-audit-online-platforms-algorithms/

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, *114*(40), 10612–10617. https://doi.org/10.1073/pnas.1706588114

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion

of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*, 331–370. https://doi.org/10.1023/A:1021240730564

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Celliers, M., & Hattingh, M. (2020). A systematic review on fake news themes reported in literature. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Responsible design, implementation and use of information and communication technology proceedings part II* (Vol. 12067, pp. 223–234). Springer International Publishing. https://doi.org/10.1007/978-3-030-45002-1_19

Chellappa, R. K., & Sin, R. G. (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management*, *6*, 181–202. https://doi.org/10.1007/s10799-005-5879-y

Council of the European Union. (2024). *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—Analysis of the final compromise text with a view to agreement* (Proposal 5662/24). https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

Directive 2000/31/EC. (2000). *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')*. European Parliament and Council. http://data.europa.eu/eli/dir/2000/31/oj

Djeffal, C., Magrani, E., & Hitrova, C. (2021). Recommender systems and autonomy: A role for regulation of design, rights, and transparency. *Indian Journal of Law and Technology*, *17*(1), Article 3. https://doi.org/10.55496/JGDN9214

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a right to explanationn is probably not the remedy you are looking for. *Duke Law & Technology Review*, *16*, 18–84. https://doi.org/10.2139/ssrn.2972855

Ekstrand, M. D., & Kluver, D. (2021). Exploring author gender in book rating and recommendation. *User Modeling and User-Adapted Interaction*, *31*(3), 377–420. https://doi.org/10.1007/s11257-020-09284-2

Engström, E., & Strimling, P. (2020). Deep learning diffusion by infusion into preexisting technologies – Implications for users and society at large. *Technology in Society*, *63*, Article 101396. https://doi.org/10.1016/j.techsoc.2020.101396

Eskens, S. (2020). The personal information sphere: An integral approach to privacy and related information and communication rights. *Journal of the Association for Information Science and Technology*, *71*(9), 1116–1128. https://doi.org/10.1002/asi.24354

European Commission. (n.d.). *Home*. European Centre for Algorithmic Transparency. https://algorithmic-transparency.ec.europa.eu/index_en

European Commission. (2020). *White paper on artificial intelligence: A European approach to excellence and trust* (White Paper COM(2020) 65 final). https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission. (2021). *Impact assessment of the regulation on artificial intelligence* (Commission Staff Working Document SWD(2021) 84 final). https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence

European Commission. (2022a). *The strengthened Code of Practice on disinformation 2022* (pp. 1–48) [Code]. https://ec.europa.eu/newsroom/dae/redirection/document/87585

European Commission. (2022b). *Review of EU consumer law* [Review]. https://commission.europa.eu/law/law-topic/consumer-protection-law/review-eu-consumer-law_en

European Commission. (2023). *Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

European Data Protection Supervisor. (2021). *EDPS Opinion on the European Commission's proposal for a Digital Services Act* (Opinion 1/2021). https://www.edps.europa.eu/data-protection/our-work/publications/opinions/digital-services-act_en

European Parliament. (2023). *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* (2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

European Parliament. (2024). *Artificial Intelligence Act. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts* (P9_TA(2024)0138). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Forssbæck, J., & Oxelheim, L. (2014). The multifaceted concept of transparency. In J. Forssbæck & L. Oxelheim (Eds.), *The Oxford handbook of economic and institutional transparency* (pp. 2–30). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199917693.013.0001

Friedman, A., Knijnenburg, B. P., Vanhecke, K., Martens, L., & Berkovsky, S. (2015). Privacy aspects of recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 649–688). Springer. https://doi.org/10.1007/978-1-4899-7637-6_19

Gavison, R. (1980). Privacy and the limits of law. *The Yale Law Journal*, *89*(3), 421–471. https://doi.org/10.2307/795891

Gayle, D. (2021, September 14). Facebook aware of Instagram's harmful effect on teenage girls, leak reveals. *The Guardian*. https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701

Haugen, F. & Committee on Internal Market and Consumer Protection. (2022, May 18). *Discussion with Frances Haugen on the global impact of the Digital Services Act*. https://multimedia.europarl.europa.eu/en/webstreaming/imco-committee-meeting_20220518-1300-COMMITTEE-IMCO

Heald, D. (2006). Varieties of transparency. In C. Hood & D. Heald (Eds.), *Transparency: The key to better governance?* (pp. 24–43). Oxford University Press. https://doi.org/10.5871/bacad/9780197263839.003.0002

Heald, D. (2022). The uses and abuses of transparency. In E. Alloa (Ed.), *This obscure thing called transparency: Politics and aesthetics of a contemporary metaphor* (pp. 37–66). Leuven University Press,. https://doi.org/10.2307/j.ctv26dhjc9

Helberger, N., Dobber, T., & de Vreese, C. (2021). Towards unfair political practices law: Learning lessons from the regulation of unfair commercial practices for online political advertising. *Journal of Intellectual Property, Information Technology and E-Commerce Law*, *12*(3), 273–296. https://hdl.handle.net/11245.1/752f60de-f725-4281-a0e7-34bbc2037a0c

Helberger, N., Sax, M., Strycharz, J., & Micklitz, H.-W. (2022). Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy*, *45*(2), 175–200. https://doi.org/10.1007/s10603-021-09500-5

High-level Expert Group On Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI* [Report]. European Commision. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Holznagel, D. (2021). Ireland cannot do it alone. *Verfassungsblog: On Matters Constitutional*. https://doi.org/10.17176/20210427-221450-0.

Hong, S., & Kim, S. H. (2016). Political polarization on Twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777–782. https://doi.org/10.1016/j.giq.2016.04.007

Horwitz, J., & Seetharaman, D. (2020, May 26). Facebook executives shut down efforts to make the site less divisive. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jungherr, A., & Schroeder, R. (2021). Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media + Society*, *7*(1), 1–13. https://doi.org/10.1177/2056305121988928

Karim, F., Oyewande, A., Abdalla, L. F., Chaudhry Ehsanullah, R., & Khan, S. (2020). Social media use and Its connection to mental health: A systematic review. *Cureus*, *12*(6), Article e8627. https://doi.org/10.7759/cureus.8627

Keles, B., McCrae, N., & Grealish, A. (2020). A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth*, *25*(1), 79–93. https://doi.org/10.1080/02673843.2019.1590851

Knowles, B., & Richards, J. T. (2021). *The sanction of authority: Promoting public trust in AI*. arXiv. https://doi.org/10.48550/arXiv.2102.04221

Koivisto, I. (2022). *The transparency paradox*. Oxford University Press. https://doi.org/10.1093/oso/9780192855466.001.0001

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*, *8*(1),

Article 117. https://doi.org/10.1057/s41599-021-00787-w

Larsson, S. (2021). AI in the EU: Ethical guidelines as a governance tool. In A. Bakardjieva Engelbrekt, K. Leijon, A. Michalski, & L. Oxelheim (Eds.), *The European Union and the technology shift* (pp. 85–111). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-63672-2_4

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1469

Larsson, S., Jensen-Urstad, A., & Heintz, F. (2021). Notified but unaware: Third party tracking online. *Critical Analysis of Law*, *8*(1), 101–120. https://doi.org/10.33137/cal.v8i1.36282

Leerssen, P. (2021). Platform research access in Article 31 of the Digital Services Act. *Verfassungsblog: On Matters Constitutional*. https://doi.org/10.17176/20210907-214355-0

Leerssen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review*, *48*, 1–13. https://doi.org/10.1016/j.clsr.2023.105790

Lorenz, P., Perset, K., & Berryhill, J. (2023). *Initial policy considerations for generative artificial intelligence* (1; OECD Artificial Intelligence Papers, Vol. 1). OECD Publishing. https://doi.org/10.1787/fae2d1e6-en

Lührmann, A., Gastaldi, L., Hirndorf, D., & Lindberg, S. I. (Eds). (2020). *Defending democracy against illiberal challengers: A resource guide* [Guide]. Varieties of Democracy Institute/University of Gothenburg. https://www.v-dem.net/documents/21/resource_guide.pdf

Meijer, A. (2014). Transparency. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford handbook of public accountability*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0043

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*, *35*, 957–967. https://doi.org/10.1007/s00146-020-00950-y

Morozov, E. (2019, February 4). Capitalism's new clothes. *The Baffler*. https://thebaffler.com/latest/capitalisms-new-clothes-morozov

O'Reilly, M., Dogra, N., Whiteman, N., Hughes, J., Eruyar, S., & Reilly, P. (2018). Is social media bad for mental health and wellbeing? Exploring the perspectives of adolescents. *Clinical Child Psychology and Psychiatry*, *23*(4), 601–613. https://doi.org/10.1177/1359104518775154

Pariser, E. (2012). The filter bubble: How the new personalized web is changing what we read and how we think. *Choice Reviews Online*, *50*(2). https://doi.org/10.5860/CHOICE.50-0926

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Plummer, L. (2017, August 22). This is how Netflix's top-secret recommendation system works. *Wired*. https://www.wired.com/story/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like/

Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425

Regulation 2016/679. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and*

on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Parliament and Council. http://data.europa.eu/eli/reg/2016/679/oj

Regulation 2022/2065. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*. European Parliament and Council. http://data.europa.eu/eli/reg/2022/2065/oj

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879

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

Sætra, H. S. (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society*, *59*, 1–10. https://doi.org/10.1016/j.techsoc.2019.04.006

Srnicek, N. (2017). The challenges of platform capitalism: Understanding the logic of a new business model. *Juncture*, *23*(4), 254–257. https://doi.org/10.1111/newe.12023

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1410

Turton, W. (2018, March 6). How YouTube's algorithm prioritizes conspiracy theories. *Vice*. https://www.vice.com/en/article/d3w9ja/how-youtubes-algorithm-prioritizes-conspiracy-theories

van Alstyne, M., & Brynjolfsson, E. (1996). Could the internet balkanize science? *Science*, *274*(5292), 1479–1480. https://doi.org/10.1126/science.274.5292.1479

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Westin, A. F. (1967). *Privacy and freedom* (1st ed.). Atheneum Press. https://openlibrary.org/books/OL5537351M/Privacy_and_freedom

Wu, B., Cheng, W.-H., Zhang, Y., Huang, Q., Li, J., & Mei, T. (2017). Sequential prediction of social media popularity with deep temporal context networks. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3062–3068. https://doi.org/10.24963/ijcai.2017/427

Xiao, B., & Benbasat, I. (2018). An empirical examination of the influence of biased personalized product recommendations on consumers' decision making outcomes. *Decision Support Systems*, *110*, 46–57. https://doi.org/10.1016/j.dss.2018.03.005

Xie, E., Yang, Q., & Yu, S. (2021). *Cooperation and competition: Algorithmic news recommendations in China's digital news landscape* [Tow Report]. Tow Center for Digital Journalism.

Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet Policy Review*, *11*(1). https://doi.org/10.14763/2022.1.1652

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136. https://doi.org/10.1080/1369118X.2016.1186713

Zadrozny, B. (2021, October 23). 'Carol's journey': What Facebook knew about how it radicalized

users. *NBC News*. https://www.nbcnews.com/tech/tech-news/facebook-knew-radicalized-users-rcna3581

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

Published by

Alexander von Humboldt
INSTITUTE FOR INTERNET
AND SOCIETY

in cooperation with

CREATe

centre
internet
et societe

R&I IN3 Internet interdisciplinary Institute
Universitat Oberta de Catalunya

UNIVERSITY OF TARTU
Johan Skytte Institute of
Political Studies