Centre on Regulation in Europe

# CROSS-CUTTING ISSUES FOR DSA SYSTEMIC RISK MANAGEMENT: AN AGENDA FOR COOPERATION

REPORT

*July 2024*

Sally Broughton Micova
Daniel Schnurr
Andrea Calef
Bryn Enstone

As provided for in CERRE's bylaws and procedural rules from its “Transparency & Independence Policy”, all CERRE research projects and reports are completed in accordance with the strictest academic independence.

The project, within the framework of which this report has been prepared, received the support and/or input of the following CERRE member organisations: Amazon, Arcom, Booking.com, Google, Microsoft, Ofcom, Tencent. However, they bear no responsibility for the contents of this report. The views expressed in this CERRE report are attributable only to the authors in a personal capacity and not to any institution with which they are associated. In addition, they do not necessarily correspond either to those of CERRE, or of any sponsor or of members of CERRE.

info@cerre.eu – www.cerre.eu

# TABLE OF CONTENTS

# ABOUT CERRE

Providing top quality studies and dissemination activities, the Centre on Regulation in Europe (CERRE) promotes robust and consistent regulation in Europe's network and digital industries. CERRE's members are regulatory authorities and operators in those industries as well as universities.

CERRE's added value is based on:

- its original, multidisciplinary, and cross-sector approach;
- the widely acknowledged academic credentials and policy experience of its team and associated staff members;
- its scientific independence and impartiality;
- the direct relevance and timeliness of its contributions to the policy and regulatory development process applicable to network industries and the markets for their services.

CERRE's activities include contributions to the development of norms, standards, and policy recommendations related to the regulation of service providers, to the specification of market rules and to improvements in the management of infrastructure in a changing political, economic, technological, and social environment. CERRE's work also aims at clarifying the respective roles of market operators, governments, and regulatory authorities, as well as at strengthening the expertise of the latter, since in many Member States, regulators are part of a relatively recent profession.

# ABOUT THE AUTHORS

**Sally Broughton Micova** is a CERRE Academic Co-Director and an Associate Professor in Communications Policy and Politics at the University of East Anglia (UEA). She is also a member of UEA's Centre for Competition Policy. Her research focuses on media and communications policy in Europe. She completed her PhD in the Department of Media and Communications at the London School of Economics and Political Science (LSE), after which she was an LSE Teaching and Research Fellow in Media Governance and Policy and Deputy Director of the LSE Media Policy Project.

**Daniel Schnurr** is a CERRE Research Fellow and a Professor of Information Systems at the University of Regensburg, where he holds the Chair of Machine Learning and Uncertainty Quantification. Previously, he was head of the research group Data Policies at the University of Passau. He received his Ph.D. in Information Systems from the Karlsruhe Institute of Technology, where he previously studied Information Engineering and Management. Daniel Schnurr has published in leading journals in Information Systems and Economics. His current research focuses on the role of artificial intelligence for competition, privacy and data sharing in digital markets, as well as regulation of AI and the data economy.

**Andrea Calef** is a Lecturer in Economics at the School of Economics of the University of East Anglia and a research member of the Centre for Competition Policy. Over time his research interest has spanned through topics, such as banking, systemic risk and international finance, ESG, Impact and Ethical Investing, FinTech and Crypto as well as competition.

**Bryn Enstone** is a research associate at the Centre of Competition Policy at the University of East Anglia and a senior research associate within the School of Politics, Philosophy, Language and Communication studies at the University of East Anglia. He holds an MA in Reasoning from the University of Kent and an MSc in International Accountancy and Financial Management from the University of East Anglia.

# EXECUTIVE SUMMARY

In August 2023, the providers of the first 19 services to be designated as very large online platforms (VLOPs) and very large online search engines (VLOSEs) under the Digital Services Act (DSA) submitted their first assessments of systemic risks stemming from the services. While the European Commission and the service providers are understandably focused on compliance, in this paper we argue that the Digital Services Board, researchers, civil society groups, and others can be looking across services and risk areas to understand what the most pressing sources of systemic risk are, where common vulnerabilities arise, and what mitigations effectively reduce negative effects.

Systemic risk assessment is an annual requirement, and we can expect a cycle of continual learning leading, ideally, to improvements in the prevention of harm and adaptation to changes in services and user behaviour. This paper sets out issues that we find cut across individual designated services and, we argue, that should be key areas for analysis across services, especially ones of similar type, as part of that process.

This paper therefore speaks to the Commission and the Board, as regulators with their overarching view and information gathering powers, and also **lays out a research agenda for academics and others who intend to contribute to the management of systemic risk** in practice and to wider debates about the governance of digital services.

Three layers can be distinguished to organise the identification and management of system risks.

1. The **service ecosystem** of a VLOP or VLOSE including the service users and its complementors.
2. **The integrated ecosystems of VLOPs and VLOSEs under common ownership**. Next to the common organisational affiliation and goals, leadership team, and shareholders, these services often share resources, insight, infrastructures, and data.
3. **The wider interconnected digital services landscape**, where VLOP and VLOSE ecosystems and other digital services are connected through shared resources, such as the use of common third-party services, and multi-homing users and complementors.

It is important to consider these different system layers in risk assessments, as harm can arise and become systemic in each of these layers but may also diffuse through these layers and become systemic in a wider sense. In this sense, holistic management of systemic risks and understanding whether risk management is sufficient, requires both an individual risk assessment per VLOP or VLOSE and a type of meta-assessment that considers the higher system layer.

The designation threshold of 45 million monthly users or 10% of the EU population set out in the DSA is a simply a minimum, and according to user numbers declared as of 31 January 2024, six designated services have user numbers more than 5 times the threshold. The potential effects of unmitigated risk and of effective mitigation for **services used by as much as half the population of the EU could be significant and merit specific attention**. At the same time the designated services differ considerably in the type of service they provide, their relationships with their users, and their business models. In this paper, we identify seven non-exclusive types in which services share some characteristics: app stores; online resources; online retail; adult services; search engines; social networks and video-sharing platforms. Nearly all these services (95%) generate revenues from advertising but the extent to which their business model depends on this varies significantly as 70% of them have at least one other revenue source. There is much to be understood about the implications of the different revenue streams and business models on risk. We therefore suggest that it would be beneficial to conduct third level analysis of collections of services **by scale, by type, and by revenue stream or business model**.

VLOPs and VLOSEs can draw from different types of governance mechanisms when orchestrating their ecosystems to ensure platform service quality and to align the interests of users and complementors to their own objectives. Such governance mechanisms can be implemented through **explicit control mechanisms or by means of incentives** that encourage user and complementor behaviour that is in the interest of the service. As these governance mechanisms are key instruments through which digital services mitigate systemic risks **there is a need to understand what works for different types of services, users, and risks,** possibly to identify best practices and certainly to enable improvement and adaptation to changes over time.

We elaborate considerations in relation to five common factors relevant to the management of risk by at least some, if not all designated services. Based on this we recommend:

1. An inclusive, cooperative process should take place, possibly led by the Digital Services Board and the Commission to set priorities among the risk areas for meta-analysis, taxonomy of harms, and strategies for consistent use of information-gathering tools across services and over time, including setting and periodically reviewing standards for the relevant metrics and data.

2. For each specific category of systemic risk set out in Article 34 of the DSA, meta-analysis across risk assessments should aim to harmonize the definitions of core concepts relevant to the risk area and the negative effects to be prevented, understandings of norms and policy goals, and data gathering and reporting.

3. The following specific areas for meta-analysis across services by independent researchers and Digital Services Coordinators through coordinated use of data access and publicly reported information including in risk assessment reports and the transparency database:
    a. Advertising business models: effects of targeting; effectiveness of ad libraries;
    b. Temporality features: possible correlation with malign use;
    c. Use of automated cross-posting tools;
    d. Very large influencers and related mitigations;
    e. Effectiveness of control and incentive mechanisms, and combinations thereof, on specific sources of risk;
    f. Recommender systems: transparency and effects of ranking signals and algorithmic curation decisions on user behaviour and collective outcomes;

   g. Inauthentic use and generative AI;
   h. Data sharing, data agglomeration, and common critical technical vulnerabilities;
   i. The roles of users, third parties and common resources or assets in content moderation.

4. As exemplified by our analysis, the meta-analysis of potential mitigation strategies to identify best practices (possibly within the categories identified in this report), which could make use of existing data but also may require experimentation, should include:
   a. An evaluation of the effectiveness of extensions to engagement-based recommender systems and algorithmic curation systems, specifically user control, explicit preference elicitation, and bridging-based algorithms.
   b. An examination of the benefits and vulnerabilities from decentralisation especially of content moderation and governance of user behaviour, and the outcome of various balances and mixes between decentralised and centralised governance mechanisms.
   c. Particular attention to aversion risk and user flight or changes in user of servicer in response to mitigations in the wider interconnected digital services landscape.

5. Building on existing cooperation mechanisms for particular types of harm (CSAM, terrorist content, disinformation, hate speech), identify opportunities where additional cooperation for risk mitigation between service providers and stakeholders is needed. For instance, through rapid response mechanisms, codes of conduct, incident databases, and coordinated intelligence gathering and mitigation approaches to safeguard against influential malign users.

# 1. INTRODUCTION

In August 2023, the providers of the first 19 services to be designated as very large online platforms (VLOPs) and very large online search engines (VLOSEs) under the Digital Services Act (DSA) submitted their first assessments of systemic risks stemming from the services. These assessments covered systemic risks from the dissemination of illegal content and systemic risk of negative effects in several areas from individual fundamental rights and health to societal systems such as civic discourse, elections, public health, and security. According to Articles 34 and 37 of the DSA, versions of these assessments along with the audit reports that should accompany them are due to be made public in autumn of 2024. Under the DSA, the European Commission is the regulatory authority for these very large services and the full risk assessments are seen only by the Commission and the Digital Service Coordinator (DSC) in the country of origin of the service provider. Therefore, the public versions of these assessments will be the first chance for the other DSCs, civil society groups, academics, and others to see how VLOP and VLOSE providers have assessed systemic risks.

Naturally, the utmost concern of providers and the Commission is compliance – understanding whether the service providers have done what the law expects of them. Hence, the focus will be on the individual assessment reports for each VLOP and VLOSE service. While the Commission has already initiated several investigations and requests for more information, the analysis and mitigation of *systemic* risks call for a wider community of stakeholders and researchers, with, we argue, involvement of the Digital Services Board (DSB), which is made up of the regulators that have direct relationships with vetted researchers and represents all the Member States, to be taking a more comprehensive view. It requires looking across services and risk areas to understand what the most pressing sources of systemic risk are, where common vulnerabilities arise, and what mitigations effectively reduce negative effects.

This paper sets out issues that cut across individual designated services and, we argue, that should be key areas for analysis. To this end, we consider categories within which cross-service analysis might be most useful and some common factors in risk assessment and prevention of harm, including several of the elements that Article 34 of the DSA instructs service providers to consider in their risk assessment, such as the design of recommender systems, content moderation systems, and data-related practices. As highlighted in this paper, these cross-cutting issues and most importantly the associated mitigation measures require further research to better understand how systemic risks emerge, spread, and can be effectively mitigated, and to critically assess the potential and the limits of the risk management approach of the DSA. Therefore, this paper lays out a research agenda for academics, civil society, and others who intend to contribute directly to the assessment of systemic risk in practice and to wider debates about the governance of digital services.

A key requirement of the risk assessment procedure under the DSA is that – much like a financial audit – it is an annually repeated process. Therefore, there is an opportunity for iterative continual learning. Consequently, improvement in quality in both the assessment and mitigation of risk is now on the table. To support such continuous improvements, the paper emphasises that sources of risks and associated mitigation measures need to be tracked and evaluated over time. Hence, the paper recommends that risk assessments should include the reporting of consistent metrics and data over time that allow for successive assessment cycles.

In addition to making recommendations for the future improvement of the risk assessment process, our contribution makes specific suggestions for the Commission, the Digital Services Board, and industry and civil society stakeholders, including VLOP and VLOSE providers, for areas of potential 'meta-analysis' across the risk assessments. Such meta-analysis can reveal the implications of each risk assessment to the wider digital services landscape and the systemic risk implications of the interactions between designated services and other digital services. Firstly, this is warranted by the DSA's notion of systemic risks, which emphasises that the risks and negative effects under consideration have system-wide implications. Secondly, the report showcases interdependencies that exist between the designated services (and also other services), which imply common vulnerabilities but also offer opportunities for risk mitigation across services ecosystems. Finally, in the context of the highlighted iterative learning process, a meta-analysis across risk assessments can facilitate and speed up learning by allowing for more generalizable insights and by establishing best practices for services of similar service types and for specific categories of systemic risks. The paper suggests that under the architecture set up by the DSA, the Commission and Digital Services Board should serve as the responsible coordinating institutions for implementing the meta-analysis of systemic risk. Taking into consideration existing mechanisms for coordination and exchange, such as those established by the Code of Practice on Disinformation and the Global Internet Forum to Counter Terrorism, they can lead a cooperative approach that will require the participation of the providers of VLOPs and VLOSEs, regulators, researchers, and various stakeholders.

Given that this paper deals with cross-cutting issues, it is important to note that the analysis here can only be viewed as a first step in identifying common factors in risk assessment and prevention of harm in the context of the DSA. Therefore, the paper will often resort to exemplifying expositions without claiming exhaustiveness. As highlighted by the analysis in this paper, the services designated as VLOPs and VLOSE are highly diverse in terms of their service types, business models, and the risks associated with them because of the broad scope of the DSA. In consequence, the analysis in this overview paper can also not be exhaustive in terms of accounting for all the specific characteristics of different services. Hence, there may exist specific exceptions to some of the arguments made here even if they are not mentioned explicitly. Despite these limitations, we believe that the analysis of some of the most pressing issues can already provide generalizable insights and provide a fruitful basis for extending the analysis to further issues and mitigation measures. In particular, the analysis in this paper can provide a blueprint for more detailed evidence gathering and analyses of specific types of systemic risks, as exemplified by the companion CERRE report on risks to electoral processes.[1]

[1] See Broughton Micova, S., & Schnurr, D. (2024). Systemic Risk in Digital Services: Benchmarks for Evaluating the Management of Risks to Electoral Processes. CERRE Report. https://cerre.eu/publications/systemic-risk-in-digital-services-benchmarks-for-evaluating-the-management-of-risks-to-electoral-processes/

The remainder of this paper is organised as follows. Section 2 discusses the peculiar characteristics of online platforms and their surrounding ecosystems. Furthermore, the services designated as VLOPs and VLOSEs are analysed and categorised with respect to their scale, service type, and business model. This emphasises the considerable heterogeneity among services but also points to shared characteristics among subsets of services. Section 3 provides a brief overview of the governance mechanisms available to very large digital services when orchestrating their surrounding service ecosystems, which are different from the conventional command-and-control mechanisms of hierarchical businesses. These governance mechanisms are of importance as they themselves may contain relevant elements to consider in the assessment of systemic risks and because they indicate what types of measures platform providers implement to mitigate systemic risks. Section 4 then analyses common factors in risk assessment and prevention of harm, including the features of content on digital services, content moderation, multi-homing, data sharing and data agglomeration, and recommender systems and algorithmic curation. Finally, Section 5 concludes and derives the main policy recommendations.

# 2. CHARACTERISTICS OF VERY LARGE DIGITAL SERVICES

## 2.1 Ecosystem Dynamics

As is typical for large digital services, each VLOP and VLOSE is at the centre of what is referred to in the literature as an ecosystem and each of these ecosystems is interwoven into a wider network of ecosystems.[2] In contrast to hierarchy-based value systems traditionally implemented by firms, large digital services (and online platforms in particular) benefit from ecosystem-based value creation by serving as intermediaries between users that retain a significant degree of autonomy.[3] This is enabled by a modular service architecture, where different actors in the ecosystem can specialise on different parts of the value creation process. For example, on a video-sharing platform, content creators can draw from different sets of individual expertise and specialise in different topics. On an e-commerce platform, merchants can independently choose their product offerings and set the prices for their products. On a search engine, advertisers can target their advertising to specific search queries in order to gain the attention of users with matching interests. As a consequence of a more open ecosystem approach, services can offer their users and complementors greater incentives for value creation and promote innovation and creativity.[4]

At the same time, the ecosystem structure implies that services cannot resort to a direct 'command-and-control' approach to influence the behaviour of their users and complementors. Here, the term 'complementors' can refer to any third party in the ecosystem that provides content, services, or products complementary to the service's offering and that together with the service create value for other users on the service.[5] For example, complementors may thus refer to app developers on an app store, to merchants on an e-commerce marketplace, or to influencers on a video-sharing platform. Instead of hierarchical 'command-and-control' mechanisms, services can draw from different types of *governance mechanisms* when orchestrating their ecosystem to ensure service quality and to align the interests of users and complementors to their own objectives.[6] Such governance mechanisms can be implemented through explicit control mechanisms (e.g., by screening, sanctioning and ultimately banning users on video-sharing platforms, or banning certain advertisers) but can also be implicitly integrated into the technical design of a service (e.g., by specifying criteria for prominent ranking or

[2] The DSA makes a distinction between very large online platforms and very large online search engines, and search engines do have a distinct intermediary function that involves indexing web sites and responding to user queries rather than hosting content on behalf of users. However, as services engaged in advertising including predictive and targeted advertising, their ecosystems are developed in that direction with advertising related complementors. The literature does not make a distinction.

[3] Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, *39*(8), 2255-2276. See also Art. 3 (i) and (j) of the DSA, which defines online platforms and online search engines as types of intermediary services based on Art. 3 (g), respectively.

[4] Gawer, A., & Cusumano, M. A. (2014). Industry platforms and ecosystem innovation. *Journal of product innovation management*, *31*(3), 417-433; Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, *39*(8), 2255-2276.

[5] Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. Strategic Management Journal, 39(8), 2255-2276. In contrast to consumers, complementors are often professional or business users, who have some kind of commercial motivation. However, in some cases, also consumers may create complementary goods on a platform, as for example in the case of Wikipedia, and can thus assume the role of a complementor.

[6] Tiwana, A. (2013). Platform Ecosystems: Aligning Architecture, Governance, and Strategy. Morgan Kaufmann.

'shadow-banning' of content). As direct control is often difficult to exert in ecosystems, services may also govern by means of incentives that encourage user and complementor behaviour that is in the interest of the service. Therefore, governance mechanisms are key instruments for digital services to address and mitigate systemic risks that could emerge or impact their ecosystem. We discuss this further in Section 3.

VLOPs and VLOSEs have, by definition, established extensive service ecosystems with very large numbers of users and complementors. A key feature of these ecosystems is the complementarities, interdependencies, and diverse relationships among the users and complementors.[7] Ergo, the ecosystem of each VLOP or VLOSE constitutes a complex system that can be prone to the emergence and propagation of systemic risks such as those detailed in Article 34 (1) of the DSA. In this ecosystem, the orchestrating service represents the central core in a hub and spoke network of interdependent stakeholders.[8] This special network topology is key to consider when assessing the sources of risks and their potential spread and contagion dynamics within these ecosystems and equally the role designated services should play in risk mitigation.

However, as highlighted in a previous CERRE report by Broughton Micova and Calef,[9] digital service ecosystems are not isolated bubbles. Instead, different ecosystems are connected through common ownership, shared resources, functionalities enabling interaction, and/or multi-homing users and complementors. Hence, systemic risks can also arise from, and propagate through, the connections and interdependencies between different service ecosystems. This includes the ecosystems of the different VLOPs and VLOSEs but also relationships with smaller online platforms, other digital services and other services such as media, advertising agencies, or data brokers.

The concept of systemic risk in digital ecosystems varies considerably from finance where the term systemic risk was first developed.[10] In finance, systemic risk typically forms and propagates through financial systems after systemic events. Whilst this can be the case in digital service ecosystems, systemic risk can also develop through the accumulation of minor harms across a large number of users and/or over extended periods of time, even on a single service.

Another facet of systemic risk in digital services worth considering is that strong interdependencies between service ecosystems exist - especially for those under common ownership. For example, the European Commission has designated Google Search, Google Maps, Google Play, Google Shopping, and YouTube under the DSA,[11] which all are owned by Alphabet through Google. Similarly, the designated VLOPs Facebook and Instagram are both owned by Meta. To benefit from economies of

[7] Broughton Micova and Calef (2023). Elements for effective systemic risk assessment under the DSA. CERRE Report. https://cerre.eu/wp-content/uploads/2023/07/CERRE-DSA-Systemic-Risk-Report.pdf

[8] Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, *39*(8), 2255-2276.

[9] Broughton Micova and Calef (2023). Elements for effective systemic risk assessment under the DSA. CERRE Report. https://cerre.eu/wp-content/uploads/2023/07/CERRE-DSA-Systemic-Risk-Report.pdf

[10] Broughton Micova and Calef (2023). Elements for effective systemic risk assessment under the DSA. CERRE Report. https://cerre.eu/wp-content/uploads/2023/07/CERRE-DSA-Systemic-Risk-Report.pdf

[11] European Commission (2024). Supervision of the designated very large online platforms and search engines under DSA. https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses

scope, services under common ownership have an economic incentive to share technical infrastructures and software components. Moreover, sharing data across services can increase the accuracy of personalized recommendations, reduce transaction costs for users, or allow for more effective targeting of ads.[12]

Shared resources have a dual implication for the management of systemic risks, which may be more likely if platforms are under common ownership. On the one hand, centralisation and shared resources can lead to common vulnerabilities, which can render the operations of services prone to the same sources of technical failure or targeted attacks. In extreme cases, this could lead to service outages or security vulnerabilities that are not confined to a single service[13] and which could be exploited by malign actors to introduce or contribute to systemic risks. For example, common security vulnerabilities across services could be exploited resulting in an increased risk to users' privacy. Malign actors could carry out concerted attacks on the operations of multiple services with the goal to spread disinformation or harmful content. On the other hand, sharing resources can scale a services' capabilities in relation to specific harms and facilitate risk mitigation measures, as countermeasures can be implemented, deployed and scaled relatively quickly compared to in more decentralised systems.[14]

Based on the above discussion, three layers of systems can be distinguished to organise the identification and management of system risks.

1. The **service ecosystem of a VLOP or VLOSE** including the service users and its complementors.
2. **The integrated ecosystems of VLOPs and VLOSEs under common ownership**. Next to the common organisational affiliation and goals, leadership team, and shareholders, these services often share resources, insight, infrastructures, and data amongst each other.
3. **The wider interconnected digital services landscape**, where VLOP/VLOSE ecosystems and other digital services are connected through shared resources, such as the use of common third-party services, and multi-homing users.

It is important to consider these different system layers in risk assessments, as harm can arise and become systemic in each of these layers but may also diffuse through these layers and become systemic in a wider sense. For instance, harm could spread bottom-up from a single service ecosystem and into the wider interconnected digital services landscape. Similarly, risk mitigation may require

[12] Krämer J., Schnurr, D. & Broughton Micova, S. (2020). The Role of Data for Digital Markets Contestability-Case Studies and Data Access Remedies. CERRE Report. https://cerre.eu/wp-content/uploads/2020/08/cerre-the_role_of_data_for_digital_markets_contestability_case_studies_and_data_access_remedies-september2020.pdf

[13] Vulnerabilities in the popular logging software Log4j or the backdoor found in the software tool "XZ utils" are just two recent examples of security vulnerabilities that affected a large share of online services with possibly critical implications for the security and data protection of these services. As the "XZ utils" incident demonstrates, popular open-source software components may be deliberately targeted by maligned actors to introduce security vulnerabilities. See, e.g., ENISA (2021). Joint Statement on Log4Shell. https://www.enisa.europa.eu/news/enisa-news/statement-on-log4shell; National Cyber Security Centre (2022). Log4j vulnerability - what everyone needs to know. https://www.ncsc.gov.uk/information/log4j-vulnerability-what-everyone-needs-to-know; Akamai Security Intelligence Group. XZ Utils Backdoor — Everything You Need to Know, and What You Can Do. https://www.akamai.com/blog/security-research/critical-linux-backdoor-xz-utils-discovered-what-to-know

[14] See also the comparison of centralised and decentralised architectures in Section 3 and on shared resources for content moderation in 4.2.

mitigation at more than one layer (e.g., beyond the single service ecosystem of a VLOP) to minimise harm, contain the spread of risks onto other layers, and avoid the risk of harm simply being shifted to other designated (or indeed non-designated) services. Our collective understanding of these dynamics and therefore ability to set priorities and guiderails for proportional mitigation is hampered by an evidence gap on these dynamics. Though the review of research in a previous CERRE report on systemic risk,[15] found evidence of cross-service contagion of certain harmful content and user behaviour, the body of evidence was also found to be very limited in relation to the harms and services in the scope of the DSA. In this sense, holistic management of systemic risk should include, in addition to the essential individual service level assessments, a type of meta-risk assessment that considers the higher system layer together with the manifold interconnections and interdependencies among service ecosystems and the wider digital services landscape.

## 2.2 Types of Service and Business Models

As of 26 May 2024, the Commission had designated 23 services as very large services. The list at that time consisted of 21 VLOPs[16] and 2 VLOSEs.[17] The list is likely to expand as other services grow and reach the threshold or as data indicating that additional services meet the criteria becomes available. Therefore, this list may be longer by the time of reading. There is already great variety among the 23 services designated as of the end of May 2024. Risk assessment and mitigation must be unique to each service due to their specific designs and functionalities, but there may be common vulnerabilities or contributions to risk across services that are of the same general type or that have similar business models. At the same time there may be lessons to learn about the effectiveness of mitigation measures by examining multiple services within those categories. Firstly, while all designated services exceed the designation threshold and have, on average, over 45 million EU users per month (approximately 10% of the EU population),[18] most of the designated services are significantly larger than this and some vastly exceed the designation threshold. Secondly, we propose seven categories for the different types of services based on the value proposition to users, the nature of the intermediation they provide, and other characteristics. Thirdly, overlapping the distinctions that can be made based on service type, we identify the main revenue streams and business models together with their respective implications for risks and risk mitigation.

[15] Broughton Micova and Calef supra note 7.

[16] As of May 2024: AliExpress, Amazon Store, The Apple App Store, Pornhub, Booking.Com, Google Play, Google Maps, Google Shopping, YouTube, Shein, LinkedIn, Facebook, Instagram, Facebook, Instagram, Pinterest, Snapchat, Stripchat, TikTok, X, Xvideos, Wikipedia and Zalando. See European Commission (2024). Supervision of the designated very large online platforms and search engines under DSA. Available at https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses.

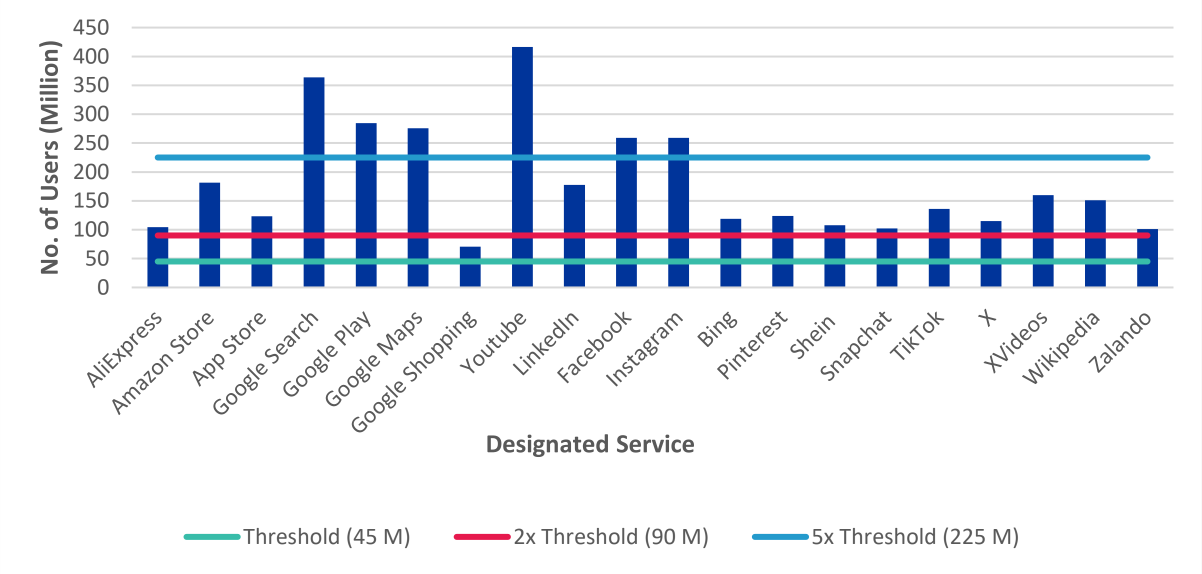
[17] Currently: Google Search and Bing Search. See European Commission (2024). Supervision of the designated very large online platforms and search engines under DSA. Available at https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses.

[18] Calculated based on declaration of user numbers by providers and according to the guidelines on determining user numbers set by the Commission in January 2023 https://ec.europa.eu/newsroom/dae/redirection/document/93451.

### 2.2.1 The scale of designated services

Figure 2.1 illustrates the scale of the 20 designated services for which data was available as of 31 January 2024. One VLOSE and five VLOPs were at least five times the designation threshold. In fact, only one designated service was not at least twice the threshold.

*Figure 2.1 Declared user numbers of designated services as of 31 January 2024*



*Source: The authors based on information on the EC's List of designated services*

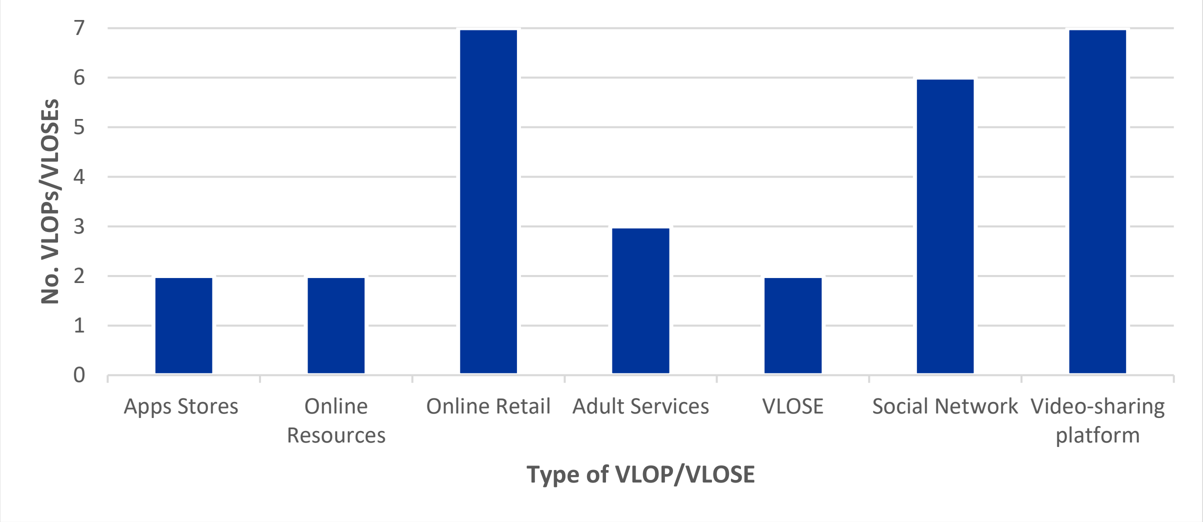
Most, if not all the designated services are large enough that harm could originate within one ecosystem and have a significant impact on society even before or without spreading to other services. Yet, there is a difference in scale between just over 90 million users and more than 225 million users in terms of the exposure of users to risks from these services. Depending on the type of service and other characteristics, this could mean that the potential accumulation of minor harms across multiple users, or the role in specific instances such as elections or public health emergencies is a more relevant consideration for risk assessments in those with such large reach than in others. Additional levels of risk may come from the sheer number of people exposed to harmful content or algorithmic bias. Those in the largest category are used by well over half of the population of the EU (448 million). The scale of all designated services also implies that potential data breaches and security failures of these services could contribute to systemic risks (especially for users' privacy; see also Section 4.3). Additional risks may come from the significance of those services as sources of information and as public spaces for expression and discourse.

In reverse, the extreme scale of these services also means that the impact of effective mitigation measures could be significant. Especially, interventions by those in the largest category have the potential to have society-shaping effects and therefore merit careful independent monitoring.

### 2.2.2 Types of service

As Figure 2.2 illustrates, the 23 services designated VLOPs and VLOSEs as of 26 May 2024 fall into several different types of services. Of the 21 VLOPs, ten are also designated video-sharing platforms (VSPs) as defined by the Audiovisual Media Services Directive (Directive 2018/1808). Six of these would also qualify as social networks, referred to in the DSA's recital 13 as a type of online platform, because of the way they are used for disseminating user-generated content across networks of users and rely on these network effects in their business model. Of the ten VSPs, three are adult services. We have maintained these as a separate category from the other VSPs because of the high-risk nature of the content they carry and the specific rules that apply to them. Another seven VLOPs are online retail services providing intermediation for the sale of products. Two VLOPs are app stores, while two others can be considered online resources - one is an online encyclopaedia, and the other is a maps application. The remaining two designated services are VLOSEs - search engines which, in response to user queries, provide algorithmically curated results based on the user's search query, their ranking criteria, and their search index.

*Figure 2.2 Types of designated services as of May 2024 (non-exclusive)*



*Source: The authors*

The way, and the extent to which different types of harm can manifest in services that enable user to user engagement varies significantly. Services differ in the type of value they offer users and complementors and the nature of the intermediation they provide, but also because their ecosystems fundamentally differ in structure. For example, some services are *'two-sided,'* connecting buyers and sellers, and deriving value from both sides. Many services are multisided, relying on network effects among and between users and complementors, for example, advertisers, app developers, data brokers, content creators, sellers, or others.[19] The extent to which the type of service relies on network

[19] Andrei Hagiu and Julian Wright, 'Multi-Sided Platforms', *International Journal of Industrial Organization* 43 (2015): 162–74; Juan Montero and Matthias Finger, 'Regulating Digital Platforms as the New Network Industries', *Competition and Regulation in Network Industries* 22, no. 2 (1 June 2021): 111–26, https://doi.org/10.1177/17835917211028787.

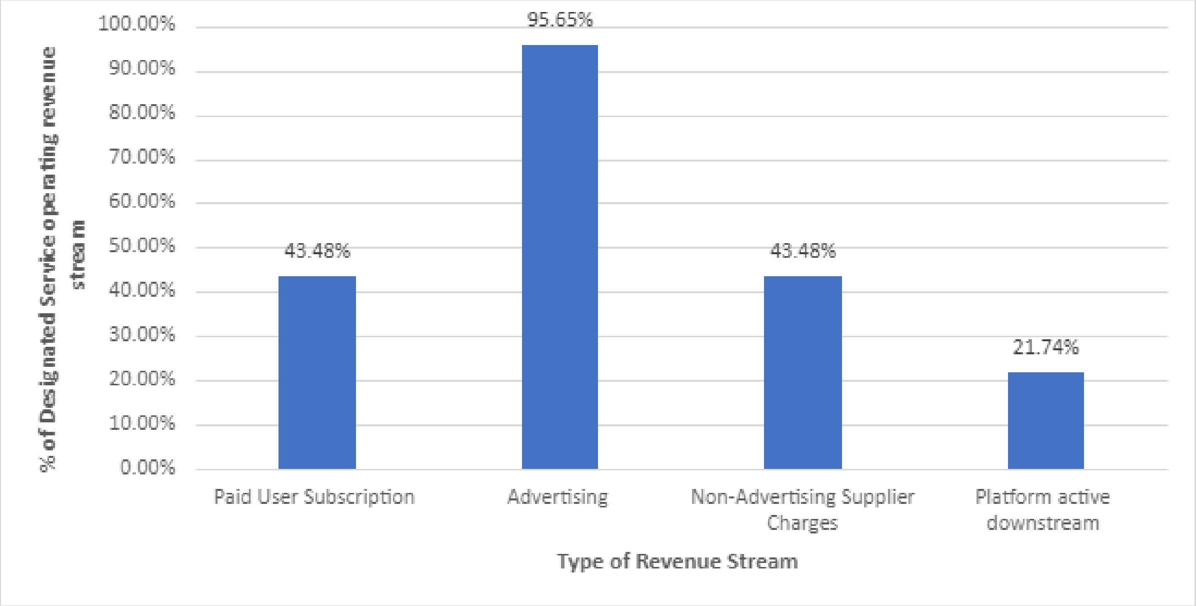
effects can have implications for the possible propagation of harm and the types of mitigation measures required.

In addition, these services are not only governed by the DSA. Other legislation may apply to specific types, which has common implications for risks associated with those types of services. For example, VSPs are regulated at the member state level in their country of origin in line with the Audio-visual Media Services Directive, which binds them to qualitative standards for advertising and the protection of users from specific types of harm, especially minors. The General Data Protection Regulation (GDPR) applies to all services, but the right to be forgotten, or de-listed, is particularly relevant for search engines.

### 2.2.3 Revenue streams and business models

As Figure 2.3 illustrates, over 95% (22/23) of designated services as of 26 May 2024 generated revenues by selling some form of advertising.[20] The extent to which their revenue stream depends on advertising varies as at least nearly 70% of those services (16/23) also had at least one additional revenue stream. These additional revenue streams include: paid-for user subscriptions, which at the time of writing were offered by at least 43% (10/23) of designated services; non-advertising based supplier charges (for example, an online marketplace charging suppliers commission or monthly fixed fees), which were offered by at least 43% (10/23) of designated services; and, almost 22% (5/23) of designated services were active downstream from their own platform (for example online market places selling their own brand products in competition with complementors).

*Figure 2.3 Percentage of designated services operating types of revenue stream*



*Source: The authors*

[20] The exception here is Wikipedia which is not operated on a for profit basis and as such as no form of advertising.

There is much still to be understood about the implications of the different revenue models for risk assessment and mitigation. The various incentive structures stemming from the way designated services generate revenue may shape their approaches to risk mitigation or the effectiveness of particular measures. For those highly dependent on advertising revenues, incentives may discourage them from deviating from engagement-driven recommender systems or employing measures that limit the propagation of content or the ability of users to expand their networks. At the same time, services that rely on advertisers' expenditure are also susceptible to the disciplining effects of major advertisers and the pressure to protect brands by effectively dealing with harmful or objectionable content.[21]

There may also be disciplining-type effects from users of subscription services that contribute to the mitigation of risk, but the subscription environment could also contribute to an increase of risk depending on the type of service and characteristics of the user base. The incentives to maintain relationships with sellers on online retail platforms or other complementors where revenues from service charges or commissions are important revenue streams may likewise shape levels of risk and approaches to mitigation. Some of these revenue streams are more dependent than others on the processing of personal data. Given that negative effects on the fundamental right to privacy are among those covered by the DSA's risk assessment and mitigation provisions, the variations in the intensity of personal data use across the different revenue streams would also define distinctions among the services. The sharing of data across services and associated protection implications are discussed further in Section 4.4.

## 2.3 Areas for Cross-service Examination

Risk assessments are conducted for each service as per the requirements of Article 34 DSA. Therefore, data provided will cover the service ecosystem level. Those service providers that own more than one service may already conduct some analysis across their designated services. Independent examination across the risk assessments of services linked together by common ownership can contribute to holding those providers accountable for assessing and mitigating risks that might stem from those interlinkages. The third layer, where VLOPs or VLOSEs are situated in the wider interconnected digital services landscape is where cross-service analysis with the involvement of a wider community of investigation involving DSCs, vetted researchers, and civil society groups is arguably most important.

We argue that it would be beneficial to examine the risk management of collections of services in groups for cross-service analysis in the third layer in three different ways:

- **By scale:** The threshold for designation as VLOPs and VLOSEs in the DSA is justifiably set at a level that allows for the designation of a wider set of services that have a far-reaching and notable impact. However, as shown above, there are some services that exceed this threshold massively and could be considered thoroughly ingrained in European society. Thinking of

[21] Rachel Griffin, 'From Brand Safety to Suitability: Advertisers in Platform Governance', *Internet Policy Review* 12, no. 3 (2023), https://policyreview.info/articles/analysis/safety-to-suitability-advertisers-in-platform-governance.

designated services just in terms of the minimum threshold therefore understates the potential of these services for systemic harm, for example when more than half the population are regular users or in some member states close to 90%. Although the categories of twice and five times the threshold we used here may not necessarily be the right categories, we argue that there is a case for looking specifically at those services that are vastly more pervasive than the minimum threshold would imply. In addition, when assessing risks and interdependencies with other services, the analysis should consider how risks depend on and may be magnified by the size of the associated service.

- **By type:** We categorise the 23 services designated as VLOPs and VLOSEs into seven non-mutually exclusive groups (app store, online resource, online retail, adult service, VLOSE, social media network, and video-sharing platform). The designated services are all unique, not only because their ecosystems may differ in the type of value they offer users and the nature of the intermediation they provide, but also because these ecosystems may fundamentally differ in structure. However, there are some similarities among services that are of the same type, which means cross-service analysis could be used to detect common sources of risk and mitigation best practices and to enable coordination in combatting malign actors or to develop tools for collaboration with vetted researchers and civil society or for engaging with users, and more.
- **By revenue stream and business model:** Despite 22 of the 23 designated services we covered here offering some form of advertising, at least 14 had some other form of revenue stream. These included 10 that offered some form of paid user subscription, 10 that offered some form of non-advertising-based supplier charges, and 5 that were active downstream from their platform. These different revenue streams each come with varying use of user data, of reliance on third parties, on levels of user autonomy and control over their experience, and other characteristics that have implications for the risk areas set out in the DSA. Eventually, regulators, the Commission and the Digital Services Board, and others will need to understand whether the risk management approach under the DSA is adequate to each of the risk areas or whether in some cases other action is needed.

# 3. GOVERNANCE MECHANISMS OF VERY LARGE DIGITAL SERVICES

This section presents the two main ways in which large digital services can exercise governance within their ecosystem. Governance in this context is commonly understood as "mechanisms through which a platform owner exerts influence over [complementors and users] participating in a platform's ecosystem."[22] A core component of governance is represented by what services refer to as "Trust and Safety” measures, as they facilitate positive relationship among users and support the forecast and detection of threats.[23] Being profit-maximisers, VLOP and VLOSE providers will implement their own platform governance but will do so primarily with their own interests at heart, i.e., to enhance their business strategy and achieve financial goals, although regulation can and does affect these choices too.

In the context of preventing and mitigating potential/actual harm, profit maximisation and harm mitigation may not necessarily always be in concert. In this sense, the DSA can be perceived as an attempt to go beyond profit-maximising incentives to internalise the potential negative externalities generated by the behaviour and outcomes on very large digital services. For example, part of the financial value produced by these services is from the interactions among the users and complementors that they enable.[24] However, various types of interactions may not be neutral or positive from the perspective of the harms and negative effects in the risk areas considered by the DSA. With respect to governing such interactions, reputation risk may be a strong incentive for service providers to prevent harm, but it may not always be sufficient to exceed the financial risks from intervening into these interactions.

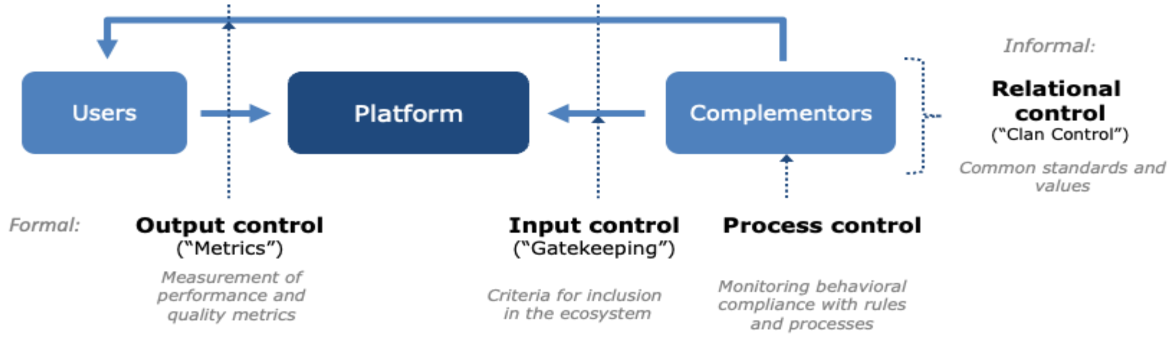
Figure 3.1 shows a stylised service ecosystem and how governance mechanisms are directed across the two main types of users any large digital service usually interacts with: complementors and end users. The figure illustrates that governance mechanisms, here specifically control mechanisms (as further discussed in Section 3.2), can target different parties and relationships in the ecosystem as well as rely on different instruments to align the behaviour of complementors and users with their own interests. Typically, digital services do not rely only on a single governance mechanism but combine approaches to orchestrate their service ecosystem.

[22] Tiwana, A. (2014). Platform Ecosystems: Aligning Architecture, Governance, and Strategy. Morgan Kaufmann. p.118.

[23] Denyer Willis, G. (2023). ‘Trust and safety’: exchange, protection and the digital market–fortress in platform capitalism. *Socio-Economic Review*, *21*(4), 1877-1895.

[24] Trabucchi, D., Muzellec, L., Ronteau, S., & Buganza, T. (2022). The platforms’ DNA: Drivers of value creation in digital two-sided platforms. *Technology Analysis & Strategic Management*, *34*(8), 891-904.

*Figure 3.1 Platform ecosystem governance flow*


Users
Platform
Complementors
Informal:
Relational control
("Clan Control")
Common standards and values
Formal:
Output control
("Metrics")
Measurement of performance and quality metrics
Input control
("Gatekeeping")
Criteria for inclusion in the ecosystem
Process control
Monitoring behavioral compliance with rules and processes

*Source: The authors*

Platforms interact with various types of users. In this chapter, especially in the part on incentives, we mainly focus on those who in the literature are referred to as complementors, although for simplicity we will often only refer to users.[25] As mentioned in Section 2.1, complementors are often business users or professional users with some kind of commercial motivation. Digital services can leverage upon two fulcrums to achieve their performance objectives when orchestrating their ecosystems:[26] i) *control*, i.e., through limiting what users can do; and ii) *incentives*, i.e., encouraging good behaviours in a variety of ways. In this vein, these two governance mechanisms could form the basis of various measures for the mitigation of risk and prevention of harm. The next two subsections investigate each of these separately.

## 3.1 Incentives

Governance mechanisms implemented via incentives are informed by the so-called 'Theory of Incentives', which considers how the design of incentives and their associated mechanisms shape the behaviour of agents, i.e., in this context, mainly complementors and users on the platform. In the context of this study, users' behaviour to achieve risk mitigation can be affected by monetary incentives, or non-monetary incentives.[27] While those are implemented in different manners depending on each digital service's business model, incentives can take various forms, such as those listed in Table 3.1.

[25] Brandenburger and Nalebuff (1996) and Carst and Yu (2023) define complementors as "downstream actors whose output enhances the value of a focal product or service that customers generate from its use". In order words, they are a segment of digital platforms' users generating values for themselves and their hosting platforms via the interactions they have with users. Full references: Brandenburger A. M., Nalebuff B. J. (1996) Co-opetition: 1 A revolutionary mindset that combines competition and cooperation 2. The Game Theory strategy that's changing the game of business, USA Carst, A. E., & Hu, Y. (2023). Complementors as ecosystem actors: a systematic review. *Management Review Quarterly*, 1-57.

[26] Chen, L., Tong, T. W., Tang, S., & Han, N. (2022). Governance and design of digital platforms: A review and future research directions on a meta-organization. *Journal of Management*, *48*(1), 147-184.

[27] [4] Aridor, G., Jiménez-Durán, R., Levy, R. E., & Song, L. (2024). The Economics of Social Media.

*Table 3.1 Incentive governance mechanisms*

| | GOVERNANCE MECHANISM | DEFINITIONS | SELECTED DESIGN FEATURES |
|---|---|---|---|
| INCENTIVE | Sharing of resources | Sharing of resources with complementors that can assist the latter in their value-creating activities | API, code library, reference design, SDK, etc. |
| | Provision of information | Provide complementors with interface- or customer-related information | Developer conferences, workshops<br>Communication channels with and between complementors and users |
| | Conferring autonomy | The extent to which digital service owners confer to complementors autonomy in conducting value-creating activities | Decentralisation of decision rights<br>Modularity |
| | Giving rewards | Giving pecuniary and nonpecuniary rewards to complementors | Revenue sharing schemes<br>Fee-based features<br>Recommendation, certification, featuring |

*Source: The authors mostly based on Chen et al. (2022).*

We briefly describe each of the governance mechanisms via incentive (listed in Table 3.1) in the following and point to potential implications and interdependencies with systemic risks as understood by the DSA. Existing studies typically consider only individual mechanisms in isolation,[28] therefore there is a need for more evidence on the interdependencies and the joint application of these mechanisms.

While the *sharing of resources* with various types of users, especially with complementors, may play a substantial role in creating/co-creating value-adding activities, it does not generally involve cash flows. This suggests that, whilst the sharing of resources within a service ecosystem has economic value, this value may not always be immediately transparent and quantifiable. Some of these

[28] Staub, N., Haki, K., Aier, S., & Winter, R. (2022). Governance mechanisms in digital platform ecosystems: addressing the generativity-control tension. Communications of the Association for Information Systems, 51(1), 43.

resources, such as APIs,[29] code libraries, reference designs and Software Development Kits (SDKs),[30] also enable users to generate insight and content, which can have positive and negative implications in the context of systemic risks and their mitigation.

As it is typically the orchestrating service that decides which resources are shared within the ecosystem, the responsibility should fall upon designated services to ensure that they share resources in a way that is congruent with the reduction of systemic risk from a harm perspective. To this end, designated services should give careful consideration to ensure that access to their services' resources or tools aimed to incentivise complementors are adequately designed with safeguards to prevent the systemic spread of harm. They should further ensure that the sharing of resources is transparently documented in the risk assessments (and their public versions) so this can also be assessed by other stakeholders, such as civil society organisations (CSOs) and researchers. It is also worth highlighting that while the design of resource sharing mechanisms may exhibit common elements across various digital services, one should not overlook service-specific elements. This specificity adds further to the rationale that the sharing of resources should be disclosed within the individual risk assessments.

Another type of incentive that digital services can implement is the *provision of information* – this can take the form of interface-related and/or customer/user-related information – to complementors to better enable them in their value-creating activities.[31] While some designated services have been very successful in attracting a large number of complementors, the initial adoption phase can be rather challenging for new complementors in terms of understanding the various functionalities of the service as a platform itself (e.g., terms and conditions, technical aspects). Thus, services are used to providing complementors with interface-related information. This can be shared in several ways, such as technical documentation, conferences, workshops,[32] or hackathons.[33] Additionally, as the features and functionalities of digital services are constantly changing over time, they are likely to provide interface-related information to complementors on a continual basis.

However, VLOPs and VLOSEs may also share customer-, interface-, or content-related information with complementors. This may involve the participation of end users, and digital services as intermediaries may facilitate the communication between complementors and users.[34] When providing either interface- or customer-related information to complementors, a service alters the information structure and the overall level of available information in its ecosystem. In fact, providing

[29] Please see Tiwana (2015a) for more details related to the impact of the introduction of APIs on developers. Full reference: Tiwana, A. (2015). Platform desertion by app developers. Journal of Management Information Systems, 32(4), 40-77.
For Apple-specific case study, please refer to Eaton et al. (2015). Full reference: Eaton, B., Elaluf-Calderwood, S., Sørensen, C., & Yoo, Y. (2015). Distributed tuning of boundary resources. MIS quarterly, 39(1), 217-244.

[30] For example, Alphabet holding provides Adroid Studio and SDKs to complementors to facilitate their activities. See Ye and Kankanhalli (2018). Full reference: Ye, H., & Kankanhalli, A. (2018). User Service Innovation on Mobile Phone Platforms. MIS quarterly, 42(1), 165-A9

[31] Schilling, M. A. (2000). Toward a general modular systems theory and its application to interfirm product modularity. Academy of management review, 25(2), 312-334.

[32] Foerderer, J. (2020). Interfirm exchange and innovation in platform ecosystems: Evidence from Apple's Worldwide Developers Conference. Management Science, 66(10), 4772-4787.

[33] Fang, T. P., Wu, A., & Clough, D. R. (2021). Platform diffusion at temporary gatherings: Social coordination and ecosystem emergence. Strategic Management Journal, 42(2), 233-272.

[34] Tan, X., Wang, Y., & Tan, Y. (2019). Impact of live chat on purchase in electronic markets: The moderating role of information cues. Information Systems Research, 30(4), 1248-1271.

interface-related information tends to increase the number of complementors, while providing customer-related information facilitates the rise in the number of interactions among existing users and complementors by reducing asymmetric information. When information is shared within an ecosystem as a way to incentivise certain behaviour or engagement by complementors, this raises questions about the role of this information in risk mitigation. Similar to resource sharing, the disclosure of such information provision should thus be included in risk assessments.

Digital services can also provide different *levels of autonomy* to complementors to enable them generating value-adding activities (including transactions with final users). An important lever is the (partial) devolution of decision rights[35] that allow complementors to more freely choose their business strategies and optimise them in a less constrained manner. Nonetheless, as this provides complementors with more freedom, it may lead to some adverse consequences in terms of less effective control and more difficult monitoring of complementors' behaviour for providers of VLOPs and VLOSEs.

Another way to incentivise complementors through autonomy is to allow them to decouple some layers of the service by effectively enabling these layers to be designed independently. In general, this is achieved by a modular architecture, i.e. by developing a system that is composed of modular subsystems. Complementors typically benefit from such modular architectures as they allow for more specialisation and differentiation, which can thus also promote overall economic efficiency and innovation. However, if decentralisation is excessive in consequence of a highly modular architecture, this can also hinder risk management activities carried out by service providers, as interventions cannot be implemented as straightforward as for a more centralised approach relying on a more monolithic and uniform system architecture.

Finally, services can offer and provide *rewards* to complementors. This is probably the most intuitive, straightforward, and widely studied mechanism to affect users' incentives. In general, such rewards can take both a pecuniary form as well as a non-pecuniary one. The former can be achieved by revenue-sharing schemes,[36] in favour of complementors or by allowing them to promote fee-generating features.[37] Conversely, the latter can be implemented through the release of certifications, recommendations, and even by featuring complementors' content or product.[38] Among the potential risks of offering rewards to complementors, a key danger is to unintentionally finance malicious users if screening activities are not well implemented or to incentivise behaviour that could contribute to systemic risks such as from the sharing of divisive content, misinformation, or hate speech. For

[35] Boudreau, K. (2010). Open platform strategies and innovation: Granting access vs. devolving control. Management science, 56(10), 1849-1872. Hagiu, A., & Wright, J. (2019). Controlling vs. enabling. Management Science, 65(2), 577-595.

[36] Miric, M., Boudreau, K. J., & Jeppesen, L. B. (2019). Protecting their digital assets: The use of formal & informal appropriability strategies by App developers. *Research Policy*, *48*(8), 103738.; Shi, X., Li, F., & Chumnumpan, P. (2021). Platform development: Emerging insights from a nascent industry. *Journal of Management*, *47*(8), 2037-2073.

[37] Claussen, J., Kretschmer, T., & Mayrhofer, P. (2013). The effects of rewarding user engagement: The case of Facebook apps. *Information Systems Research*, *24*(1), 186-200.

[38] Rietveld, J., Schilling, M. A., & Bellavitis, C. (2019). Platform strategy: Managing ecosystem value through selective promotion of complements. *Organization Science*, *30*(6), 1232-1251.

example, product recommendations on an e-commerce platform may unintendedly support sellers of counterfeit products.

## 3.2 Control

Governance mechanisms can also be implemented via a 'control-based' approach. Among the large variety of control mechanisms, we are going to briefly present and analyse those listed in Table 3.2, as they are most likely to have implications for DSA risk management purposes. As illustrated by Figure 3.1 above, control mechanisms may target different pathways, relationships, or parties in a service ecosystem, which has implications on how easily they can be scaled or how much information is necessary to effectively implement these different mechanisms. While the academic literature has studied control mechanisms often in relation to complementors, they have the advantage over incentive-based mechanisms in that they can also influence and sanction the behaviour of users, whose ecosystem participation and service use are driven much less by economic considerations. Therefore, control mechanisms seem particularly suited to address the behaviour of malign users, for whom economic incentives are less relevant.

*Table 3.2 Control governance mechanisms*

| | GOVERNANCE MECHANISM | DEFINITIONS | SELECTED DESIGN FEATURES |
|---|---|---|---|
| CONTROL | Access control / input control | Governance mechanisms determining who is allowed to join the platform and use digital interfaces | Screening mechanisms<br>Restriction on the use of boundary resources<br>Access fees |
| | Output control | Evaluation and monitoring of complementors' outputs and outcomes | Reputation scores, online reviews, ratings, etc |
| | Behavioural control / process control | Deciding on the types of interactions allowed or deemed appropriate on the platform | Anti-manipulation techniques<br>Restriction on the exchange of contact information |
| | External relationship control | The extent to which digital service providers allow complementors to interact with other platforms | Exclusive relationships<br>Reduction of compatibility |

*Source: The authors mostly based on Chen et al. (2022).*

The control of who can join (*access*) the platform and benefit from the use of its features as well as the interactions with the users within the platform ecosystem is probably the best known and most popular control mechanism. VLOSEs have an incentive to be as inclusive as possible in the search results they return and VLOPs have a general incentive to attract a large number of users and complementors. However, VLOPs also have the privilege, responsibility, and incentive to exclude specific complementors, users, or content created (e.g., to ensure the service's quality or to comply with regulatory obligations). In this way, these companies can orchestrate their service ecosystems and align users' interests with their own interests by screening users[39] and taking action against them if required to mitigate risks. Platform interests can be shaped by a variety of factors including commercial incentives, commitments in self-regulatory mechanism, regulation, reputational concerns, and others.

A restriction of access can be achieved by the imposition of fees to be paid by complementors,[40] which will typically result in a rationing of the supply of these complementors. Digital services can also restrict the access to the use of boundary resources (such as through APIs and SDKs), which determine the feasible actions and put limits on the possible behaviour of complementors and (to some extent) users when interacting via the service.[41]

Contemporaneously or alternatively, providers of VLOPs and VLOSEs can focus on monitoring *output* and outcomes provided by complementors. This can be achieved via user feedback and reputation systems,[42] accumulated reputation scores,[43] and more generally online ratings and reviews.[44]

Digital services can also directly influence complementors' behaviours by setting rules and establishing processes that determine the allowed types of interactions (*behaviour control*). For example, they can set up so-called anti-manipulation mechanisms for online reviews[45] that may eventually result in the suspension of accounts[46] or the reduction of information provided to a complementor,[47] even though the latter could have negative effects for a VLOP itself.[48] In general, this type of control is relatively

[39] Tiwana, A. (2015). Evolutionary competition in platform ecosystems. *Information Systems Research*, *26*(2), 266-281; Song, P., Xue, L., Rai, A., & Zhang, C. (2015). The ecosystem of software platform: A study of asymmetric cross-side network effects and platform governance. *Available at SSRN 2568817*.

[40] Hossain, T., Minor, D., & Morgan, J. (2011). Competing matchmakers: an experimental analysis. *Management Science*, *57*(11), 1913-1925; Dushnitsky, G., Piva, E., & Rossi-Lamastra, C. (2022). Investigating the mix of strategic choices and performance of transaction platforms: Evidence from the crowdfunding setting. *Strategic Management Journal*, *43*(3), 563-598.

[41] Gawer, A. (2021). Digital platforms' boundaries: The interplay of firm scope, platform sides, and digital interfaces. *Long Range Planning*, *54*(5), 102045.

[42] Bolton, G., Greiner, B., & Ockenfels, A. (2013). Engineering trust: reciprocity in the production of reputation information. *Management science*, *59*(2), 265-285.

[43] Li, J., Chen, L., Yi, J., Mao, J., & Liao, J. (2019). Ecosystem-specific advantages in international digital commerce. *Journal of International Business Studies*, *50*, 1448-1463; Fan, Y., Ju, J., & Xiao, M. (2016). Reputation premium and reputation management: Evidence from the largest e-commerce platform in China. *International Journal of Industrial Organization*, *46*, 63-76.

[44] Choi, A. A., Cho, D., Yim, D., Moon, J. Y., & Oh, W. 2019. When seeing helps believing: The interactive effects of previews and reviews on e-book purchases. Information Systems Research, 30(4): 1164-1183.

[45] Kumar, N., Venugopal, D., Qiu, L., & Kumar, S. (2018). Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems*, *35*(1), 350-380.

[46] Reischauer, G. & Mair, J. (2018). How Organizations Strategically Govern Online Communities: Lessons from the Sharing Economy. *Academy of Management Discoveries* 4(3). 220–247.

[47] Zhu, F., & Marco Iansiti, M. Why Some Platforms Thrive and Others Don't. *Harvard Business Review* 97(1), 118–125.

[48] Gu, G., & Zhu, F. (2021). Trust and disintermediation: Evidence from an online freelance marketplace. *Management Science*, *67*(2), 794-807.

resource-intensive from the perspective of a service as it requires continuous monitoring of user behaviour. In consequence, service providers have often resorted to algorithmic approaches that can automate monitoring and control (see, e.g., content moderation, which is further discussed in Section 4.3).

## 3.3 Implications of Governance Mechanisms for Mitigation of Systemic Risks

Large digital services can exercise governance over complementors and users via the means of various types of incentives and control mechanisms. In order to fulfil the DSA's requirements, each VLOP or VLOSE provider will have chosen the optimal mix in relation to its own business strategy. Therefore, it is to be expected that the mix of governance mechanisms does not only differ among designated services and will change over time, but also that this mix will and most likely should also differ across DSA risk areas.

Governance mechanisms, especially those relating to incentives, can be associated with risks themselves. For example, conferring autonomy of decisions or sharing resources, such as giving access to critical APIs, could be exploited by malign users. Moreover, conferring greater autonomy to complementors could present the risk that the effectiveness of control mechanisms, and hence eventually measures to mitigate systemic risks, could be undermined. On the contrary, excessive control could also lead to systemic risks such as limiting legitimate contributions to civic discourse. Therefore, our analysis above considered some of the possible contributions of selected mechanisms to systemic risks. Hence, we derive some insights on the implication of each governance mechanism per se, although an exhaustive analysis will require a more extensive in-depth investigation.

In this context, it is important to acknowledge that, as mentioned, most large digital services do not only use several governance mechanisms, including both incentives and control mechanisms, simultaneously, but that each sevice has adopted its peculiar and evolving service-specific configuration mix of governance mechanisms. Therefore, it is important that risk assessments by providers of VLOPs/VLOSEs consider the effects of individual governance mechanisms and assess the overall short-term and long-term impact generated by the combination of their specific mix of governance mechanisms in an aggregate manner.

Such a high level of complexity requires the development of sophisticated techniques to determine whether systemic risk will ultimately be reduced in a significant manner by the governance mechanisms put in place by each VLOP or VLOSE. As the current level of comprehension of such phenomenon is relatively limited, further and stronger collaborations between large digital services and independent parties, such as civil society organisations, academic and non-academic researchers are not only desirable, but needed to assess the effectiveness (and, if there are, limitations) of these mechanisms in mitigating systemic risks.

# 4. COMMON FACTORS IN RISK ASSESSMENT AND PREVENTION OF HARM

As shown in the previous sections, the services designated as VLOPs and VLOSEs are all very different. Though they can be grouped into some categories according to type and business model, as we have done, they still have distinct features, functionalities, and designs. Nevertheless, there are some elements that are common to groups of them and significant enough to the potential for risk and assessment of successful mitigation that they merit cross-platform examination and independent analysis. In this section, we deal first with the features of content then multi-homing by users who disseminate content in particular through automated means or with significant reach. We then consider factors on the services' side beginning with the sharing of data by and across designated services particularly those under common ownership. We then address considerations in relation to content moderation and finally, we cover recommender systems and algorithmic curation, which in different ways are used in nearly all designated services. This can include product recommendations, social media feeds, or search query results.

## 4.1 Content Features

All VLOPs and VLOSEs handle content of some form. It could be a user-uploaded video or image, an advertisement, an image of a product for sale, a review left by a customer, or a host of other types and forms. Amidst this great variety of content, we argue here that there are four aspects of content characteristics that merit cross-service analysis to better understand how they contribute to risk and how they are accounted for in mitigation measures. These are content format, temporality, whether or how content is regulated at source, and whether content is AI-generated.

### 4.1.1 Content Format

Since television came into people's homes it has been assumed that audiovisual content is particularly impactful on audiences, particularly when *pushed* at audiences rather than chosen, or *pulled*, as one would buy a newspaper and select specific parts to read. Early national and eventually European laws, such as the Television without Frontiers Directive, reflected these understandings of the power of audiovisual media.[49] Though there remain debates within communities of media effects researchers, there is evidence to support the idea that audiovisual content is particularly powerful, especially on children, which has long raised concerns about the effects of audiovisual advertising.[50] VLOPs in particular are used for sharing audiovisual content and have been developing mitigation measures adapted to that format of content. There is sufficient evidence on the power of this format to merit cross-cutting analysis of the effectiveness of these measures and the sharing of best practices across services. At the same time, as the now famous case of disinformation through a deepfake circulated

[49] Natali Helberger, 'From Eyeball to Creator-Toying with Audience Empowerment in the Audiovisual Media Service Directive', *Entertainment Law Review*, 2008; Sally Broughton Micova, 'From the Television without Frontiers Directive to the Audiovisual Media Services Directive', in *Audiovisual Policy in Motion*, ed. Hertiana Ranaivoson, Sally Broughton Micova, and Tim Raats (Routledge, 2023).

[50] For an overview of evidence from audience and psychology research see J.K. Maher, M.Y. Hu, and R.H. Kolbe, 'Children's Recall of Television Ad Elements: An Examination of Audiovisual Effects', *Journal of Advertising* 35, no. 1 (2006): 23–33, https://doi.org/10.2753/JOA0091-3367350102.

in the 2023 Slovakian election demonstrated, the audio-only format should not be forgotten in mitigation measures. Another aspect to examine across risk assessment reports is the extent to which mitigation measures are being designed to deal with all the content formats that can be disseminated.

Content formats can also be thought of in terms of the genre that it explicitly displays or hypertextually elicits. Content that is advertising may use a format viewers will recognise as a drama with a narrative and characters or may adopt the expert endorsement format. User-generated content may also elicit drama, news reporting, comedic, or other recognisable formats. There is significant evidence from research and observation of disinformation online that news-like formats or 'fake news' are a particular source of concern and can take the form of websites that may appear in search results as well as be shared by users of VLOPs. As Wardle and Derakhshan point out in their seminal work for the Council of Europe, there are several reasons for this.[51] The news format evokes a culturally defined, ritualistic relationship with its audience, and assumptions about itself as a trusted source of information. At the same time, the pace of news cycles and the draw of novelty in those cycles contribute to the potential for disinformation to be convincing and/or to transfer to legitimate news sources. Those designated services that tend to host or index news format content would therefore merit particular attention in cross-cutting analysis. Most of these are also parties to the Code of Practice on Disinformation, which already provides for information sharing and exchange on best practices over mitigation.

### 4.1.2 Temporality

There are two elements to consider in relation to the temporality aspect of content on VLOPs in particular, the level of permanence and how immediately consumers consume this content, i.e., whether it is uploaded or live-streamed. At the time of writing, the issue of temporality relates mainly to the functionalities afforded by VLOPs that host user-generated content, however, as services evolve this might change. An examination of how these service providers have considered the degree of risk related to the temporal characteristics of content, and how they have attempted to mitigate the risks would produce relevant insight as to the level of risk inherent in some common platform functionalities and challenges for mitigation.

On some services, content is transient and appears to users only briefly, which may limit the harm it can cause but also may limit the effectiveness of user flagging mechanisms and other mitigation measures, as well as the ability of public authorities and civil society actors to detect problems. On other services, content, once uploaded or shared, stays until removed by the user or the service provider. Content moderation, as discussed further in Section 4.2, is crucial to risk mitigation on such services complemented by user flagging or reporting. The potential for propagation of such lasting content is greater, especially if searchable and public. Indeed, much of the research that has evidenced

[51] Wardle, C. and Derakhshan, H. (2017) Information disorder: Toward an interdisciplinary framework for research and policy making, Council of Europe https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

propagation patterns of dissemination or hate speech has focused on this kind of content.[52] The implications of large quantities of old content that might pre-date some mitigation measures are not yet widely understood, though service providers may have insight specific to the services they operate.

The unique dangers of live transmission, or live-streaming, have been well-recognised in relation to risks from terrorist content, where the aim of the perpetrators is wide dissemination,[53] or in instances of child sexual abuse, where more live-streaming to narrower audiences has been implicated as an important facilitator of abuse.[54] Both of these relate to the dissemination of illegal content and the live transmission to others of criminal acts. There are already transnational coordination and cooperation mechanisms involving several of the designated services, especially those also designated as video-sharing platforms, aimed at improving the mitigation of these risks.[55] Efforts to combat these two types of illegal content have not yet been fully successful and the problem of child sex abuse material (CSAM) in particular shows no sign of abating.[56] The DSA risk assessments provide an opportunity to understand the role that existing mechanisms play in mitigation efforts and where the gaps are.

There are also risks of negative effects in relation to gender-based violence, other aspects of the protection of minors, and individuals' well-being, among others.[57] Live-streaming can be accompanied by comments and reactions in real time. Much of the research into harassment, bullying, and other forms of online violence has focused on live-streaming by gamers on specialist platforms and harassment of streamers themselves. These have highlighted the important and complex roles that community moderators play in mitigating harassment in live-stream chats and comments, including in the face of coordinated 'hate raids' that use social bots to bombard live streams.[58] However, there are also 'live' functionalities on several VLOPs and similar options to comment or react in real time in a continuous stream. Live-streaming presents specific challenges to content moderation efforts by platforms because of its immediacy and the speed at which compliance with terms or guidelines needs

[52] For example, Jinyin Chen et al., 'Research on Fake News Detection Based on Diffusion Growth Rate', ed. Lei Yu, *Wireless Communications and Mobile Computing* 2022 (14 July 2022): 6329014, https://doi.org/10.1155/2022/6329014; Lynnette Hui Xian Ng, Iain J Cruickshank, and Kathleen M Carley, 'Cross-Platform Information Spread during the January 6th Capitol Riots', *Social Network Analysis and Mining* 12, no. 1 (2022): 133.

[53] Alessia Zornetta and Ilka Pohland, 'Legal and Technical Trade-Offs in the Content Moderation of Terrorist Live-Streaming', *International Journal of Law and Information Technology* 30, no. 3 (2022): 302–20.

[54] For a recent review of the evidence see Larissa S. Christensen and Jodie Woods, '"It's Like POOF and It's Gone": The Live-Streaming of Child Sexual Abuse', *Sexuality & Culture*, 9 January 2024, https://doi.org/10.1007/s12119-023-10186-9.

[55] Sally Broughton Micova and Ivana Kostovska, 'Chapter 2 Video-Sharing Platforms', in *Study on the Implementation of the New Provisions in the Revised Audiovisual Media Services Directive (AVMSD)*, ed. European Union (Brussels: European Commission, 2021), 26–127, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=73303.

[56] Annual report of the Internet Watch Foundation 2023 https://iwf.org.uk/annual-report-2023/

[57] DSA Article 34(1)

[58] Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel, 'Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities', in *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences*, IMX '21 (New York, NY, USA: Association for Computing Machinery, 2021), 61–72, https://doi.org/10.1145/3452918.3458796; Jie Cai et al., 'Hate Raids on Twitch: Understanding Real-Time Human-Bot Coordinated Attacks in Live Streaming Communities', *Proc. ACM Hum.-Comput. Interact.* 7, no. CSCW2 (October 2023), https://doi.org/10.1145/3610191.

to be assessed and actioned. Research into the ways people circumvent platforms' rules have found that using live-streaming functionalities is one of the ways this is done.[59]

### 4.1.3 Regulated vs non-regulated content

Within the vast amounts of content that VLOPs and VLOSEs host and index respectively, a significant portion comes from institutions that are regulated or participate in self-regulatory mechanisms. These are professional media bound by broadcasting rules in member states, editorial and journalistic codes of ethics, or advertising codes. Unlike in the case of content created by other kinds of users, whether a social media post, an online video or a website, there is ostensibly harm mitigation taking place at the source. For example, the content coming from a television channel would be regulated in its country of origin, and content from a daily news publisher could be expected to comply with the code of the self-regulatory body in which it participates and arguably the wider journalistic ethics.[60] Some of the risk mitigation measures used by designated services rely extensively on this type of content, particularly ones aimed at combatting disinformation. Such content may be referred to as a source in nudging mechanisms, or prioritised by algorithms returning search queries or as part of bridging-based recommender systems because it is taken as coming from an authoritative source (see also Section 4.5.4).

However, content from sources that fall under regulatory or self-regulatory regimes is not necessarily free from risk. There are self-regulatory systems across Europe that struggle to enforce ethical standards and member states whose regulatory authorities' independence is challenged.[61] There are threats to the independence and quality of media for political and economic reasons in many member states. These issues in media environments were part of the impetus for the European Media Freedom Act (EMFA), which came into force on 7 May 2024 and contains provisions aimed at ensuring independence of both media services and regulatory authorities.

The EMFA's Article 18 requires VLOPs that host content of media service providers to set up systems in which those media service providers, including publishers, can declare themselves as such and as editorially independent and subject to regulatory requirements in a member state. VLOPs then must ensure those media services receive expedited processing of their complaints and be identified as such publicly on the platform. The EMFA also requires VLOPs to report annually on their rejection of designations of media service providers, any restrictions or suspensions on declared media services, and dialogue undertaken in cases of dispute of such actions. Data in these reports can be combined with information from systemic risk assessments and the DSA transparency database as well as other sources to generate insight into the role of regulated media content and the risks covered by the DSA. Cross-service analysis should be conducted that answers questions about the relative contribution of media services' content to negative effects and to risk mitigation measures. The Digital Services Board

[59] Rosalie Gillett, Joanne E. Gray, and D. Bondy Valdovinos Kaye, '"Just a Little Hack": Investigating Cultures of Content Moderation Circumvention by Facebook Users', *New Media & Society*, 1 February 2023, 14614448221147661, https://doi.org/10.1177/14614448221147661.

[60] Such as the International Federation of Journalists Code https://www.ifj.org/who/rules-and-policy/global-charter-of-ethics-for-journalists

[61] See recent data from the Media Pluralism Monitor https://cmpf.eui.eu/media-pluralism-monitor-2023/

can coordinate with the European Board for Media Services[62] to make use of its power to convene a structured dialogue with industry and civil society representatives.[63]

### 4.1.4 AI-generated content

The rise of generative AI, and especially the programmatic access to AI tools and the availability of open-source large-language models, presents new challenges to the assessment and mitigation of various risks. While the use of generative AI does not necessarily create new types of risks, it increases the scale of content generation. As AI-generated content is now difficult to distinguish from manually-created content, identifying the authenticity of content (both manually and programmatically) has become a challenge with obvious risk implications such as for the dissemination of hate speech and misinformation (see, e.g., the ongoing discourse about 'deep fakes' and their potential impact on the digital information space). By now, there have been numerous examples of generative AI being used to spread harmful content,[64] while there is also emerging evidence of inherent biases in its generated content.[65] Whereas the use of generative AI on digital services by users and complementors can have positive effects (such as automated bots that assist users or new ways to generate content), inauthentic use of generative AI represents a particular concern, especially when this is exploited by malign users.

Given the rapid developments in this context and the potential large-scale impact of AI-generated content, there arises the need to swiftly develop common (adaptive) standards on what can be considered appropriate levels and types of automated content generation and on what constitutes manipulative intervention. On the one hand, this is a cross-cutting issue that affects all VLOPs and VLOSEs to some extent. On the other hand, the acceptable standard may differ between the types of systemic risks (e.g., risks to the electoral process may call for stricter standards),[66] which suggests that risk-specific standards should be developed through inclusive multi-stakeholder processes involving industry, regulators, independent domain experts, and civil society organisations.

## 4.2 Multi-Homing

Most of us are multi-homing users who have profiles and share content on multiple platforms and use online retail services. Most of us are not intentional or significant contributors to the negative effects the DSA aims to prevent. However, the non-excludable nature of these services that allow the average user to multi-home also allows bad actors to multi-home, and multi-homing across platforms gives

[62] Under the EMFA the Board replaces the European Regulators Group for Audiovisual (ERGA) which was convened by the Commission and codified in the 2018 revision of the Audiovisual Media Service Directive. EMFA's article 19 sets out the nature and scope of the structured dialogue.

[63] There are several Digital Service Coordinators that are also members of the Board for Media Services.

[64] Karhu, K., & Ritala, P. (2021). Slicing the cake without baking it: Opportunistic platform entry strategies in digital markets. *Long Range Planning*, *54*(5), 101988.

[65] Motoki, F. Y., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice*, *198*(1), 3-23; Motoki, F. Y., Pinho Neto, V., & Rangel, V. (2024). Digital Dissonance: Large Language Models' Unbalanced Political Narrative. https://ssrn.com/abstract=4773936

[66] See for a more elaborate discussion Broughton Micova, S., & Schnurr, D. (2024). Systemic Risk in Digital Services: Benchmarks for Evaluating the Management of Risks to Electoral Processes. CERRE Report. https://cerre.eu/publications/systemic-risk-in-digital-services-benchmarks-for-evaluating-the-management-of-risks-to-electoral-processes/

some content creators the ability to reach very large audiences. Therefore, it should be a consideration in the assessment and mitigation of risk for VLOPs that allow users to disseminate content. We identify two issues that can be considered significant interlinkages among services and therefore are worth cross-service examination and possibly coordination on mitigation. The first is automated cross-posting, and the second is the category of very large influencers.

### 4.2.1 Cross-posting and assisted cross-posting

Cross-posting or mass posting to multiple platforms using automated tools has been found to be less common than posts that either originate on a platform or that are manually copied across from another platform, but there is evidence that content disseminated through automated cross-posting tools are more likely to be harmful. For example, a 2018 study by Gerlitz and Rieder found that one of the more popular third-party applications that allows for the cross-posting of content across many platforms only accounted for 0.46% of posts on a popular social media site.[67] However, that same study also found that  the automated cross-posts contained a disproportionate amount of harmful content, especially sexually explicit content. Research into foreign information manipulation and interference (FIMI) has also found automated cross-posting, often using inauthentic 'aggregator accounts,' is used to amplify content, obfuscate its origins, and create illusions of authentic discussion.[68] There is evidence of similar use of automated cross-posting in other malicious disinformation campaigns as well.[69]

The average multi-homing user arguably has less incentive to cross-post, much less to learn and use automated tools for doing so. Active multi-homing and cross-posting create a burden on non-professional users,[70] who may be less willing to share the additional data and spend the time managing accounts. Commercial content creators (or aspiring ones) and users with malicious intent are more likely to be aiming to maximise their reach across multiple services. Therefore, it would be useful for the percentage of posts made using first-party and third-party cross-posting tools to be disclosed in risk assessments along their contribution to actioned content figures. More research is also needed into the use of these automated tools, especially as the proportion of posts made using them may rise in the future with the advances in AI.

### 4.2.2 Influential content creating complementors

Some users of VLOPs in the category of social media and video-sharing platforms have extremely high reach across multiple platforms. These highly influential content creators, or influencers, can be considered complementors and usually derive commercial value through their activity on the

[67] Gerlitz, C. and Rieder, B., 2018. Tweets Are Not Created Equal: Investigating Twitter's Client Ecosystem. *International Journal of Communication (19328036)*, *12*.

[68] European Union External Action Service (January 2024) 2nd EEAS Report on Foreign Information Manipulation and Interference Threats A Framework for Networked Defence. https://www.eeas.europa.eu/eeas/2nd-eeas-report-foreign-information-manipulation-and-interference-threats_en; Josephine Lukito, 'Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three US Social Media Platforms, 2015 to 2017', *Political Communication* 37, no. 2 (2020): 238–55.

[69] Tom Wilson and Kate Starbird, 'Cross-Platform Disinformation Campaigns: Lessons Learned and next Steps', *Harvard Kennedy School Misinformation Review* 1, no. 1 (2020).

[70] Montero, J. and Finger, M., 2021. Regulating digital platforms as the new network industries. *Competition and Regulation in Network Industries*, *22*(2), pp.111-126.

platform.[71] Some reach audiences that dwarf those of even national broadcast media in many member states. For example, Andrew Tate, whose profiles and content across multiple platforms were eventually removed or blocked for violations associated with the risk areas in the DSA, had follower and subscriber numbers nearing 5 million and operated multiple accounts, profiles, and channels on various services when they were actioned in 2022.[72] For comparison, the VRT, one of Belgium's public service broadcasters had claimed an audience of just over 2 million the previous year. While many of those reached by such very large influencers will be outside of the EU, the potential for harm even within the member states can be considerable.

Most very large influencers are likely benign or at least not intending to contribute to negative effects in society through their use. Many are otherwise celebrities and others are simply commercial influencers who promote products and services through their content or create content in order to benefit from ad revenue sharing opportunities offered by the VLOPs and contracts with advertisers. Arguably, commercial influencers can be governed through incentive-based mechanisms tied to their potential to monetise their content or engage in commercial communication. Many member states also have rules for influencers as part of their advertising standards self- or co-regulatory systems,[73] however, these rules have largely focused on disclosure and transparency requirements and there are myriad other concerns that relate to some of the systemic risk areas mentioned in the DSA.

Antoniou recently identified some categories not being addressed under such systems, namely 'risk-fluencing' which can be a source of negative effects to public health, minors and anyone's physical well-being, and 'stereo-scripting' which relies on harmful stereotypes,[74] with implications for fundamental rights. Services that are designated as video-sharing platforms under the AVMSD should have experience in dealing with this as they should already be taking measures to ensure that commercial communications on their services do not promote discrimination or encourage behaviour prejudicial to health and safety.[75] There may already be evidence that can be examined and best practices that can be shared in relation to effective mitigation and what (dis)incentives work with commercial influencers, and where gaps might exist.

Very large multi-homing influencers also merit attention in cross-service analysis and information sharing among service providers because there will be some who are not solely or not wholly motivated by commercial incentives and are using platforms deliberately to achieve negative effects in the areas covered by the DSA.

[71] Catalina Goanta and Sofia Ranchorás, 'The Regulation of Social Media Influencers: An Introduction', in *The Regulation of Social Media Influencers*, ed. Catalina Goanta and Sofia Ranchorás (Edward Elgar Publishing, 2020), 47–73.

[72] Wilson, J. (2022) "The Downfall Of Andrew Tate And Its Implications" Forbes.https://www.forbes.com/sites/joshwilson/2022/08/30/the-downfall-of-andrew-tate-and-its-implications/; Sung, M. (2022) "Andrew Tate banned from YouTube, TikTok, Facebook and Instagram" NBCNews https://www.nbcnews.com/pop-culture/viral/andrew-tate-facebook-instagram-ban-meta-rcna43998

[73] See discussion and examples in Sally Broughton Micova and Ivana Kostovska, 'Chapter 2 Video-Sharing Platforms', in *Study on the Implementation of the New Provisions in the Revised Audiovisual Media Services Directive (AVMSD)*, ed. European Union (Brussels: European Commission, 2021), 26–127, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=73303.

[74] Alexandros Antoniou, 'When Likes Go Rogue: Advertising Standards and the Malpractice of Unruly Social Media Influencers', *Journal of Media Law*, 2024, 1–44.

[75] See Articles 28b(2) and Article 9(1) in the AVMSD.

There is no exact threshold of followers or reach that qualifies a user as an influencer and it could be that audiences are accumulated across multiple services (e.g reaching younger audiences on one and older on another). VLOPs have varying criteria that must be met for users to engage in revenue sharing, and requirements for other commercial arrangements such sponsorship or promotion deals will be set by advertisers and their agencies. Service providers might find it useful to arrive at some common understanding of thresholds of reach that trigger concern or at least particular attention in mitigation measures. It can be that malign influencers have significant reach only on one platform but connect to specific audiences on others. Risk assessment reports should be examined across services that fall into the category of social media and video-sharing platforms to see if consideration of the reach of users blocked or the extent to which violating content is generated by single users has been considered and reported on. In addition, independent research across services to better understand the incentive structures and governance mechanisms for very large influencers would help identify issues and best practices in mitigating negative effects from malign or inadvertently negative effects.

## 4.3 Data Sharing and Data Agglomeration

All VLOPs and VLOSEs rely on data-driven services and access to manifold sources of data to some degree. Given that negative effects on the fundamental right to privacy are among those covered explicitly by the DSA's risk assessment and mitigation provisions (see Art. 34 (1) (b) DSA), the sharing and agglomeration of data represents a relevant cross-cutting issue and is therefore discussed here. At the same time, privacy risks are the main focus of the General Data Protection Regulation, and thus assessment and mitigation of privacy risks under the DSA should consider the relation with these provisions and existing risk mitigation measures developed under this legislation. Furthermore, the Digital Markets Act considers some risks from data agglomeration, even though the focus here is on harm to effective competition. Given these more specific data-related regulations, risk assessment and risk mitigation under the DSA should focus on those privacy risks that may remain beyond the scope and enforcement of these other regulations and that may emerge especially in conjunction with other risks considered under the DSA.

Sharing of user data can yield benefits for users, because such data may allow digital services to improve the quality of their service, provide more accurate personalized recommendations (see also Section 4.5), select and prioritise content that is of higher relevance to a particular user, or to better target ads to the preferences of users. At the same time, such sharing of user data across services can entail privacy risks for individuals, as the combination of data can allow data holders to infer new insights about the underlying data subject. In this vein, the sharing and combination of user data may create systemic risks according to Art. 34 (1) (b) of the DSA, even if users have given consent for the use of their personal data for a specific purpose.[76] For instance, this could be the case if comprehensive user profiles were leaked to third parties that are not authorised to access the data and/or that may

[76] There may also arise competition issues from data agglomeration and data-driven market power (see Fast, V., Schnurr, D., & Wohlfarth, M. (2023). Regulation of data-driven market power in the digital economy: Business value creation and competitive advantages from big data. *Journal of Information Technology*, *38*(2), 202-229). However, these issues are not discussed further here, as they are more likely to be assessed and dealt with under the Digital Markets Act than the Digital Services Act.

use the data to the detriment of users. A harm to privacy could also occur if combined personal data were used for purposes that go beyond the specific purpose of users' consent. Moreover, the combination of fine-granular behavioural user data can possibly reveal the identity of a data subject (and thus further personal information about that data subject) even if the user data is collected and stored in pseudonymized form.[77] In this case, active risk mitigation measures based on privacy-preserving technologies (such as k-anonymity or differential privacy)[78] are necessary to prevent systemic privacy risks.

On the other hand, shared data can facilitate the identification and mitigation of risks. For instance, the sharing of information between platforms can facilitate the detection of malicious content or behaviour (see Section 4.4). By sharing knowledge about potential risks and malign actors, platforms can therefore contribute to the safeguarding against contagion of risks across different platforms. In this context, it is important to anticipate that such shared data and the underlying technical infrastructures may become a point of attack themselves, as malign actors may deliberately target these shared resources to maximise the impact of their attacks or to systematically undermine the moderation of specific behaviour or content. Furthermore, for platforms under common ownership, the platform provider may be in a particularly advantageous position to identify risks by observing both user behaviour and content flow across platforms. At the same time, most providers of VLOPs and VLOSEs have also the necessary capabilities and infrastructure resources to analyse such user data on a larger scale, which could facilitate the identification of malign actors and harmful content.

## 4.4 Content Moderation

Content moderation is arguably the cornerstone of risk mitigation on VLOPs that enable users to share content but also relevant to those that host product reviews or other types of comments. Social media and video-sharing platforms moderate vast quantities of user-generated content and online retail services moderate product reviews and possibly question-and-answer posts or other content. There are a variety of ways this is done. Automated content moderation plays a significant role on these very large platforms, but human moderators are also used. Here we raise two issues that we argue merit investigation at a cross-service level to better understand potential vulnerabilities and the effectiveness of mitigation. The first is the implications of the various levels and types of decentralisation of the implementation of content moderation. The second is the extent to which content moderation relies on resources shared by multiple services.

[77] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*(1), 1-5. De Montjoye, Y. A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536-539. Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, *10*(1), 1-9.

[78] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, *10*(5), 557-570; Dwork, C. (2006). Differential privacy. In (M. Bugliesi, B. Preneel, V. Sassone & I. Wegener) *International Colloquium on Automata, Languages, and Programming. 33rd International Colloquium, ICALP 2006 Venice, Italy, July 10-14, 2006 Proceedings, Part II* (pp. 1-12). Springer. Berlin Heidelberg.

### 4.4.1 Decentralisation vs centralisation

Content moderation approaches and practices vary across designated services. This is largely a function of the variations in scale, service type, and business model as identified in Section 2 as well as the design features and functionalities of the services. The DSA affords a significant degree of autonomy in how VLOPs operationalise content moderation and requires them to make sure moderation is proportionate to the risks. One way in which their approaches differ is in the extent to which they decentralise content moderation, though ultimately responsibility remains with the VLOP provider.

As Jhaver et al.[79] highlight, the term decentralisation can have many meanings when applied to content moderation. Arguably, the obligations of the DSA encourage designated services to centralise their content moderation at least to some extent, while at the same time encouraging the engagement of some external actors such as fact checkers. Yet even within these relatively centralised ecosystems, decentralisation can, and does, occur in at least two ways.

*Decentralisation via outsourcing to third-party content moderators:*
There is significant evidence that designated services outsource content moderation to third parties. For example, a 2022 New York Times article found that at least 5,000 employees at Accenture were being employed for content moderation on behalf of a VLOP client.[80] Ahamad and Krzwdzinski[81] and numerous reports from journalistic investigations provide examples of content moderation for social media platforms being carried out in different countries often far from where the moderated content originated.[82]

In these instances, content moderation is still mostly top down, i.e., the content moderation policies are coming from the management of designated services but the ‘day to day’ work of content moderation, especially that done by humans, has been outsourced. Outsourcing may provide access to lingual expertise, cultural context, or technical knowledge which the designated services do not possess ‘in house’, which may contribute to more effective mitigation of risk. However, where firms are only financially incentivised to outsource to reduce their cost base, this may affect the quality and timeliness of content moderation, and consequently the level of risk. Another key issue here is that designated services may use the same external third-party content moderators, which would then constitute a resource shared among services that interlinks these services with what could be a common vulnerability (see also Section 4.2.2). As Broughton Micova and Calef argued in a previous

[79] Jhaver, S., Frey, S. and Zhang, A.X., 2023. Decentralizing platform power: A design space of multi-level governance in online social platforms. *Social Media+ Society*, *9*(4), p.20563051231207857.

[80] New York Times (2022) ‘How Facebook relies on Accenture to scrub toxic content’ How Facebook Relies on Accenture to Scrub Toxic Content - The New York Times (nytimes.com)

[81] Ahmad, S. and Krzywdzinski, M., 2022. Moderating in obscurity: how Indian content moderators work in global content moderation value chains. In *Digital work in the planetary market* (pp. 77-95). Cambridge, MA, Ottawa: The MIT Press, International Development Research Centre.

[82] Ahmad and Krzwdzinski highlight that India and the Philippines are particularly common in this regard. However, there are also media articles which report outsourcing to several different countries, for example this BBC article discusses the outsourcing of content moderation to Kenya. Firm regrets taking Facebook moderation work - BBC News

CERRE Report,[83] such interlinkages should be considered by VLOP providers in their assessment of risk. Disclosure and transparency about the use of outsourcing will be important to the effective oversight of DSA implementation, and cross-platform analysis and information sharing among services can help reduce the potential that outsourcing heightens the vulnerability of content moderation systems.

*Decentralisation via multi-level governance and user autonomy:*

Decentralisation should also be considered in terms of the multi-level governance mechanisms (especially those relating to control) inherent in the design of many VLOPs.[84] This is a common feature of social media and video-sharing platforms where active users share some responsibility for preventing harm.[85] In these cases, governance is still largely centralised, however elements of governance exist at middle levels, such as through volunteer moderators, or even at the user level, such as through filtering options or third-party-provided 'block lists' that can be chosen and applied by users.

For example, some social media services provide functionalities that allow users to create groups or communities in which those who created and/or manage the group play a moderating role in addition to the centralised content moderation done by the platform.[86] Others provide access to resources and functionalities to third parties who curate 'block-lists' of accounts to block based on certain preferences. These lists then essentially serve as filters for users who choose to apply them. Video-sharing platforms may give channel owners the ability to moderate comments on their videos and to block users from accessing their channels in addition to the platform's own measures. Wikipedia stands out as the most decentralised service, which is in line with its not-for-profit business model based on contributions. Each language edition of the service has its own processes for governing articles, which are supported by committees, developers, and operators within the Wikimedia Foundation.

As can be seen in all these examples, a degree of control over content is transferred to users and complementors playing various roles. There are obvious benefits to this. Firstly, users in these roles may have a greater contextual understanding of the circumstances under which they are applying content moderation. This is particularly relevant to content moderation under the DSA as not all the harms subject to systemic risk in the DSA are 'black and white'. Whilst it may be more straightforward to identify a post which contains sexually explicit images, it is perhaps less easy to identify which posts could lead to a negative effect on civic discourse or what constitutes gender-based violence as local or other types of contextual understanding might be needed. Secondly, decentralising content moderation in this manner necessitates a degree of transparency to function and so may reduce the opaqueness of content moderation.

[83] Supra note 2.

[84] Supra note 64.

[85] Sally Broughton Micova and Ľuboš Kukliš, 'Responsibilities of Video-Sharing Platforms and Their Users', in *Audiovisual Policy in Motion*, ed. Hertiana Ranaivoson, Sally Broughton Micova, and Tim Raats (Routledge, 2023).

[86] Jhavier, S. et al. supra note 63. Testing

However, decentralisation of content moderation is not without risk, which is why a comparison of the approaches pursued by the VLOPs could be useful. The platform is relinquishing a degree of control over content moderation, which may affect the uniformity, timelessness, and quality of content moderation. For example, volunteer moderators bring risks of judgement errors and biases, or may be malign actors infiltrating that role. Some approaches involving users may be more vulnerable to manipulation by malign users than others or may include safeguards that could be considered best practices. Because of the vast differences in the designs and functionalities of the VLOPs, this is the kind of issue that would likely be best examined within the service type categories discussed in Section 2.2.2.

### 4.4.2 Shared resources for content moderation

Some content moderation requires a degree of accessible and collaboratively maintained shared resources for content moderation. Indeed, some of these resources such as shared hash databases for identifying illegal content are now fundamental in the prevention of the dissemination of harm across digital ecosystems and the wider digital services landscape and require input from several different platforms and digital services, including designated services.

Shared resources and coordination across platforms play an important role in stopping harm from becoming systemic to the wider digital services landscape and indeed throughout society. For example, if a video of terrorist content originates on a non-designated service and spreads to a designated service where it is detected, then the designated service needs a feedback mechanism by which it can alert other services, including the originating services that the video has been shared and identified as terrorist content. The same holds if the video originates on the designated platform and is then shared to other platforms. For terrorist content such mechanisms exist under the auspices of the Global Internet Forum to Counter Terrorism (GIFCT)'s incident response system.[87] Tools (such as a hash database) through which platforms can help to take pre-emptive steps to stop the harm being shared throughout the wider digital services landscape help to mitigate risk; GIFCT's hash-sharing database and accompanying tools are good examples of these. Such shared resources for identifying content, common technical tools, and response mechanisms are inherently shared resources that cross boundaries of the individual platform ecosystems and any vulnerabilities in those are also part of those interlinkages.

Furthermore, there are other ways in which a resource might be shared. For example, within common ownership structures a content moderation tool might be shared for the sake of economic and technical efficiency. Moreover, when third parties are used in content moderation (such as fact checkers) it is possible that designated services may use the same companies (see Section 4.2.1). Fact checkers, for example that are part of international certification schemes are often used by multiple services in their efforts to combat disinformation. Equally, as AI plays an increasingly larger role in content moderation, it is possible that automated systems could be shared between designated services or that interdependencies arise from common training data for AI-based systems. Equally,

[87] https://gifct.org/

academic research has shown that AI tools may be used to measure the radicalness of users,[88] and therefore the degree to which content moderation resources are shared in the future could increase with increased automation. This is especially noteworthy, as AI-based systems may themselves be prone to contributing to systemic risks, e.g., through systematically biased predictions.

## 4.5 Recommender Systems and Algorithmic Curation

Many VLOPs use recommender systems to provide users with personalised recommendations for content, products, or other items. Especially for social media services, personalised recommendations and personalised feeds of content have become the default for selecting and distributing content to users. By establishing innovative mechanisms for content discovery, recommender systems have been among the core drivers of the business success of some VLOPs (e.g., the popularity of TikTok has commonly been attributed to its recommendation algorithm[89]). Furthermore, recommender systems create manifold business value, e.g., by increasing service providers' revenues and profits when used for product recommendations to online shoppers or when delivering content recommendations for paid media.[90] Similarly, online search engines use algorithmic ranking techniques that, among other ranking signals, order search results according to previous user interactions with the search results page.[91] In this vein, search results that receive more positive user engagement (e.g., clicks on a search result without immediate return to the search results page) are regularly considered more relevant and thus ranked higher, everything else being equal. In addition, search engines commonly personalise search results by adjusting the ranking of retrieved search results according to information about the individual user and the individual context of the search query.[92]

### 4.5.1 Automated information curation and the impact on user behaviour and collective outcomes

Given the information overload that users are exposed to in the digital sphere, recommender systems and search engines have become central information gateways that determine what information,

[88]See Sofat, C. and Bansal, D., 2023. RadScore: An automated technique to measure radicalness score of online social media users. *Cybernetics and Systems*, *54*(4), pp.406-431. RadScore: An Automated Technique to Measure Radicalness Score of Online Social Media Users (tandfonline.com)

[89] Hern, A. (2022). How TikTok's algorithm made it a success: 'It pushes the boundaries'. The Guardian. https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity

[90] Lee, D., & Hosanagar, K. (2021). How do product attributes and reviews moderate the impact of recommender systems through purchase stages?. *Management Science*, *67*(1), 524-546.; Kumar, A., & Hosanagar, K. (2019). Measuring the value of recommendation links on product demand. *Information Systems Research*, *30*(3), 819-838.; Fast, V., Schnurr, D., & Wohlfarth, M. (2023). Regulation of data-driven market power in the digital economy: Business value creation and competitive advantages from big data. *Journal of Information Technology*, *38*(2), 202-229.; Krämer, J., Schnurr, D., & Micova, S. B. (2020). *The role of data for digital markets contestability: case studies and data access remedies*. CERRE Report. https://cerre.eu/publications/data-digital-markets-contestability-case-studies-and-data-access-remedies/

[91] See, e.g., Agichtein, E., Brill, E., & Dumais, S. (2006, August). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-26); Muller, B. (n.d.) How Search Engines Work: Crawling, Indexing, and Ranking. https://moz.com/beginners-guide-to-seo/how-search-engines-operate

[92] See, e.g., Microsoft (n.d.). Privacy - Browsing and search https://account.microsoft.com/privacy/web-search; Google (n.d.). Personalization & Google Search results. https://support.google.com/websearch/answer/12410098?hl=en; Krämer, J., Schnurr, D., & Micova, S. B. (2020). *The role of data for digital markets contestability: case studies and data access remedies*. CERRE Report. https://cerre.eu/publications/data-digital-markets-contestability-case-studies-and-data-access-remedies/;

content, and products users discover, interact with, and consume.[93] An important distinction is that search engines are pull-based mechanisms that curate results in response to a search query that is initiated by the user's action, whereas recommender systems are push-based mechanisms that deliver automated recommendations based on the current context of the user and past user interactions without an explicit user query. Hence, search engines allow a user to more actively steer and control the information discovery process. In contrast, recommender systems often try to anticipate user preferences and intentions based on users' observed behaviour and without requesting users' active input. In contrast to traditional media services, information aggregators, and retailers, both recommender systems and search engines share the feature that items are not curated manually anymore but selected and ranked in an automated manner. Both mechanisms leave the user with agency to actively engage with the suggested items (i.e., to actually click on a link, interact with content, or purchase a product) and possibly to further choose among a pre-selected list of items.

However, a large body of research finds that users will often follow the suggestions of these mechanisms and demonstrates that especially the presented order of items has a significant impact on what items users will eventually consume.[94] In practice, the importance of prominence and visibility on large online platforms, search engines, devices, and app stores are illustrated by the large amounts of money that firms are willing to pay to have prioritised visibility (as, for example, illustrated by the popularity of paid search advertisements or the large compensations paid for being the pre-selected default service on devices).[95] Furthermore, the prominence on platforms has been the subject of recent competition law cases emphasising the power of some very large digital services to act as major information gateways and to influence downstream user decisions and transactions.

Therefore, besides paid advertising, the 'organic' ranking and curation of information by recommender systems and search engines have critical implications on what is seen by users in digital environments and what is eventually consumed.[96] Consequently, the criteria and algorithms that determine the selection and order of items play a crucial role for the implications that these mechanisms could have on users' access to information, the dissemination of content, and the popularity of products, apps, and services. The DSA stipulates that providers of VLOPs and VLOSEs must assess how the "design of their recommender systems and any other relevant algorithmic

[93] See, e.g., Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science*, *381*(6656), 398-404.

[94] See, e.g., Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research*, *29*(1), 84-102; Kumar, A., & Hosanagar, K. (2019). Measuring the value of recommendation links on product demand. *Information Systems Research*, *30*(3), 819-838; Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, *37*(4), 530-552; Edelman, B., & Lai, Z. (2016). Design of Search Engine Services: Channel Interdependence in Search Engine Results. Journal of Marketing Research, 53(6), 881-900 ; Boczkowski, P., Mitchelstein, E., & Matassi, M. (2017). Incidental news: How young people consume news on social media. *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*; Bandy, J., & Diakopoulos, N. (2023). Facebook's News Feed Algorithm and the 2020 US Election. *Social Media+ Society*, *9*(3), 20563051231196898.

[95] See, e.g., Feiner, L. (2023). Google paid $26 billion in 2021 to become the default search engine on browsers and phones. CNBC. https://www.cnbc.com/2023/10/27/google-paid-26-billion-in-2021-to-become-a-default-search-engine.html

[96] As described above, the distinction between pull-based mechanisms implemented by search engines and push-based mechanisms implemented by recommender systems has implications for the degree to which users' explicit intent and control shape the algorithmically curated and retrieved information. In either case, the selection and ranking of the information has an important impact on users' final consumption.

system" could affect systemic risks (Art. 34 DSA) and also refers to "testing and adapting [these] algorithmic systems" as a risk mitigation measure (Art. 35 DSA). In addition, the DSA imposes transparency obligations for the criteria that determine the selection and ranking decisions of recommender systems, and their (relative) influence (see Art. 27). Thus, the DSA aims to provide more clarity on the factors that lead to a recommendation, which is necessary to assess the consequences of such selection and ranking criteria on individual user behaviour and collective outcomes.

### 4.5.2 Ranking signals and evaluation criteria for automated information curation

The opaqueness of these criteria to external parties has made it difficult in the past for independent researchers and external organizations to evaluate the implications of these critical inputs empirically. In particular, the consequences of choosing alternative ranking criteria or applying different weightings are therefore under-researched. This is of particular importance, as there have been widespread concerns that especially providers of social media services rely too much on user engagement as the dominant benchmark to evaluate and adjust recommender systems and other algorithmic curation systems.[97] As providers of digital services, especially those that are advertising-funded (see also Section 2.2.3), have an economic incentive to maximise the time users spend on their services, there is a rationale to maximise user engagement and retention.[98] This is further exacerbated for services characterised by strong network effects, as the increased presence and contributions of one user directly or indirectly add to the utility of other users, and hence lead to additional revenue opportunities from increased service usage or higher willingness to pay.

Recommender systems generally rely on the assumption of revealed preferences, that is, users' preferences can be inferred from their actual behaviour and choices. Thus, if a user chooses between a set of proposed items, the selected item is inferred to deliver the maximum utility to this user. In reverse, if a user follows the recommendation of a recommender system this is taken as evidence that the system has proposed an accurate recommendation in the interest of the user. While the concept of revealed preferences is a common assumption, not only in the context of recommender systems but in many scientific disciplines such as economics and social sciences, this concept is not without criticism and limitations.[99] External influence (e.g., through persuasion efforts), varying contexts and suggestive environments (e.g., by the design of user interfaces), individual beliefs and bounded rationality, or adaptable expectations represent exemplary reasons why observed choices and thereby

[97] See, e.g., Bouchaud, P. (2024). Skewed perspectives: examining the influence of engagement maximization on content diversity in social media feeds. *Journal of Computational Social Science*, 1-19; The relative downsides and risks of engagement-based metrics are especially relevant for sensitive content categories such as politics, which has for example been acknowledged by Meta. See in Hagey K., Horwitz J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

[98] Cf. Persily, N. (2022). Platform Transparency: Understanding the Impact of Social Media. Testimony Before the United States Senate Committee on the Judiciary. Subcommittee on Privacy, Technology, and the Law. https://www.judiciary.senate.gov/imo/media/doc/Persily%20Testimony.pdf; Bundtzen, S. & Schweizer, C. (2023). Access to Social Media Data for Public Interest Research: Lessons Learnt and Recommendations for Strengthening Initiatives in the EU and Beyond. Institute for Strategic Dialogue. https://www.isdglobal.org/isd-publications/researcher-access-to-social-media-data-lessons-learnt-recommendations-for-strengthening-initiatives-in-the-eu-beyond/

[99] Thorburn, L., Bengani, P., & Stray, J. (2022). What does it mean to give someone what they want? the nature of preferences in recommender systems. *Understanding Recommenders*. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157

inferred preferences may not align with the true self-interest of the user.[100] Addictive behaviour and mental disorders represent extreme examples that illustrate why individuals' choices should not necessarily be equated with their true preferences and the utility-maximizing options. Instead, more careful reflection and possible extensions to the revealed preferences approach are desirable, especially if actions and choices of individuals are influenced by these systems that are considered important for their own and collective well-being. This is further supported by arguments that human preferences are not necessarily stable but are often constructed on-the-fly when faced with a particular decision situation. This has been extensively documented by psychological research and can create a range of behavioural biases (e.g., through framing of the decision situation), which call for further caution when inferring well-being and utility from users' observed choices.[101]

Furthermore, it has been found that interactions and consumption patterns that create the most short-term attention and user engagements are often those that are not conducive to a healthy and informative discourse or the dissemination of accurate and representative information.[102] In fact, there are well-documented examples that extremism, toxicity, conspiracy theories, and hate speech can attract wide user attention and generate intensive user engagement. This was acknowledged by Meta in 2018 when Mark Zuckerberg referred to a "basic incentive problem" and internal research that demonstrated that people would engage more with content as this content gets closer to the line of what is actually allowed as content on the platform.[103] Especially on services that enable dissemination of user-generated content, such as social networks and video-sharing platforms, this may not only make this type of content more visible to users due to 'viral' sharing and algorithmic amplification but also promote the creation of more divisive and 'borderline' content.[104] There exist some economic incentives for service providers to mitigate these issues when they face public backlash from their user base or when they fear churn of either users that are negatively affected by hate speech or of advertisers that are concerned about brand safety (see also Section 2.2.3). Some VLOPs claim to have undertaken efforts to demote content that is close to violations of their community policies.[105] On the contrary, other providers of VLOPs, in the name of freedom of speech, have announced plans to roll back earlier efforts to reduce the visibility of harmful content.[106]

[100] Ibid.

[101] See, e.g., Lichtenstein, S. (2006). *The Construction of Preference.* Cambridge University Press.

[102] See, e.g., Brady, W. J., Jackson, J. C., Lindström, B., & Crockett, M. J. (2023). Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, *27*(10), 947-960; Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2023). Engagement, user satisfaction, and the amplification of divisive content on social media. https://doi.org/10.48550/arXiv.2305.16941

[103] Constine, J. (2018). Facebook will change algorithm to demote "borderline content" that almost violates policies. https://techcrunch.com/2018/11/15/facebook-borderline-content/; Zuckerberg, M. (2018). A Blueprint for Content Governance and Enforcement. https://perma.cc/ZK5C-ZTSX

[104] See the citation of the statement of, J. Perreti, CEO of BuzzFeed, according to which "the most divisive content that publishers produced was going viral on the platform [Facebook], he said, creating an incentive to produce more of it" in Hagey K., Horwitz J. (2021). Facebook tried to make its platform a healthier place. It got angrier instead. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215

[105] See, e.g., Meta (n.d.). Content borderline to the Community Standards. https://transparency.meta.com/en-gb/features/approach-to-ranking/content-distribution-guidelines/content-borderline-to-the-community-standards

[106] Perez, S. (2023). Musk says X will address shadowbanning 'soon,' but former Trust & Safety exec explains why that will be difficult. https://techcrunch.com/2023/08/17/musk-says-x-will-address-shadowbanning-soon-but-former-trust-safety-exec-explains-why-that-will-be-difficult/

Finally, there is a controversial debate about whether and how much recommender systems and algorithmic curation contribute to polarisation and other adverse effects as a consequence of possibly exposing users to less diverse information and more homogeneous opinions of similar peers.[107] Such 'filter bubble' or 'echo chamber' effects could occur when personalised recommendations or search results were to reinforce users' access and consumption of one-sided information while making it less likely to see contrasting information or more diverse opinions. Whether recommender systems in particular lead to more homogeneous information consumption and whether exposure to less diverse opinions indeed have negative effects, such as increased political polarisation, are both subject to ongoing research and diverging findings, which can also be viewed as an indication that context-specific factors play an important role.[108]

### 4.5.3 Automated information curation as a potential contributing factor to systemic risks and the need for cooperation and iterative learning to develop appropriate risk mitigation measures

The above discussion shows that automated information curation can contribute to the emergence or dissemination of systemic risks, which is influenced in particular by the design and functioning of the underlying technical systems. Most notably, risks can arise if recommender systems and algorithmic curation systems focus too narrowly on engagement-based metrics and evaluation benchmarks. For a more detailed conceptualisation and systematisation of these risks, domain-specific analyses are needed, which is beyond the scope of this cross-cutting issue paper.[109] However, the above discussion indicates that conventional systems for generating personalised recommendations and algorithmic curation can be prone to risks without additional adjustments to or extensions of these technical systems. Concerns about the impact of automated information processing are further magnified by the increasing use of generative AI for retrieving and curating information, as the recommendations and results of these systems are more probabilistic and thus much more difficult to anticipate and to evaluation.[110] In addition, these AI-based systems are much more opaque and highly dependent on the underlying training data. In this context, risk assessments under the DSA should provide transparency and clarity about the measures that individual providers of VLOPs and VLOSEs have

[107] See, e.g., Terren, L., & Borge-Bravo, R. (2021). Echo Chambers on Social Media: A Systematic Review of the Literature. *Review of Communication Research*, *9*, 99-118; Lerman, K., Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118.; Nguyen, T. T., Hui, P. M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web* (pp. 677-686). Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., ... & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137-144.

[108] See, e.g., Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science*, *381*(6656), 398-404; Lerman, K., Feldman, D., He, Z., & Rao, A. (2024). Affective polarization and dynamics of information spread in online networks. *npj Complexity*, *1*(1), 8; Levy, R. E. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, *111*(3), 831-870; Keijzer, M. A., & Mäs, M. (2022). The complex link between filter bubbles and opinion polarization. *Data Science*, *5*(2), 139-166.

[109] For an analysis of risks in the context of electoral processes, see Broughton Micova, S., & Schnurr, D. (2024). Systemic Risk in Digital Services: Benchmarks for Evaluating the Management of Risks to Electoral Processes. CERRE Report. https://cerre.eu/publications/systemic-risk-in-digital-services-benchmarks-for-evaluating-the-management-of-risks-to-electoral-processes/

[110] See, e.g., Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7001-7025).

taken to identify and mitigate such risks as well as their efforts to evaluate the implications and outcomes of these measures on a continuous basis.

To date, independent researchers regularly lack access to the necessary data and/or platform services to run experimental analyses that could inform our understanding of the impact of algorithmic design decisions on systemic risks and their mitigation.[111] At the same time, providers of VLOPs and VLOSEs seem to differ considerably in their efforts to investigate the impact of ranking metrics and recommendation criteria on user behaviour and collective outcomes as well as the underlying causal mechanisms. Even in cases where such research is undertaken, the economic incentives often advise against releasing the findings of this research publicly. Here, the DSA could serve as an instrument to overcome an inherent incentive problem: service providers may not voluntarily start to research and analyse potential risks, as knowledge about potential negative effects would compel them to act upon this information, possibly requiring them to sacrifice business interests. Recent cases suggest that ignoring a known problem may indeed generate worse public backlash than not undertaking the research at all.[112]

In this vein, the DSA could facilitate cooperative research efforts between providers of VLOPs and VLOSEs and independent researchers aimed at deriving generalizable insights on how the technical design of systems and algorithms could contribute to the minimization and mitigation of systemic risks. For this purpose, risk assessments under the DSA can serve as an instrument to coordinate and continually develop these research efforts by involving the Digital Services Board as a central authority, while also ensuring accountability. To this end, risk assessments should ensure that research endeavours are executed according to pre-defined parameters and processes as well as that findings and outcomes are reported transparently and consistently across assessment cycles according to pre-specified metrics.[113] The periodic risk assessments could therefore help to learn about how recommender systems and algorithmic curation contribute to the emergence and proliferation of systemic risks as well as mechanisms and approaches to mitigate them both within and across ecosystems as well as at the wider digital services landscape layer.

### 4.5.4 Possible risk mitigation measures in the context of automated information processing

A process of iterative learning and continuous evaluation is necessary as there are no easy universal answers to addressing the concerns about engagement-based recommendations and algorithmic curation, which have so far been most prominently voiced in the context of very large social network services. While there is a myriad of proposals on how recommender systems and algorithmic curation

[111] Cf. Persily, N. (2022). Platform Transparency: Understanding the Impact of Social Media. Testimony Before the United States Senate Committee on the Judiciary. Subcommittee on Privacy, Technology, and the Law. https://www.judiciary.senate.gov/imo/media/doc/Persily%20Testimony.pdf; Bundtzen, S. & Schweizer, C. (2023). Access to Social Media Data for Public Interest Research: Lessons Learnt and Recommendations for Strengthening Initiatives in the EU and Beyond. Institute for Strategic Dialogue. https://www.isdglobal.org/isd-publications/researcher-access-to-social-media-data-lessons-learnt-recommendations-for-strengthening-initiatives-in-the-eu-beyond/

[112] See, e.g., the reporting about internal research at Facebook in a series of articles by The Wall Street Journal on the “the facebook files” https://www.wsj.com/articles/the-facebook-files-11631713039 and the associated public reaction.

[113] Cf. Stray, J. (2021). Designing recommender systems to depolarize. https://doi.org/10.48550/arXiv.2107.04953

systems could be improved upon, there is a comparable number of objections towards these proposals and their potential limitations. Hence, experimental investigations and empirical evaluations in the actual platform settings are necessary to assess the effectiveness of the various proposed design changes at mitigating systemic risks.

Approaches for improving on the concept of revealed preferences and user engagement as the major primary criterion for recommendation systems and algorithmic curation include but are not limited to the following options:[114]

### *Individual-level intervention: User controls and nudges*

One mitigation strategy to address potential risks from automated information curation is to provide users with more control and increase their agency over the selection, curation, and filtering of information presented to them.[115] This can include the integration of additional technical tools offered by the platform itself or developed by third parties (see, e.g., the proposal of a decentralised middleware layer[116] that would offer consumers the ability to choose from a variety of third-party-provided information curation mechanisms). Furthermore, the use of digital nudges and boosts has been proposed to foster a choice architecture that can help users make more deliberate decisions about their information gathering and content consumption that are less prone to be influenced by behavioural biases and contextual influence.[117]

The DSA itself imposes a mandatory risk mitigation measure aimed at more user control over recommender systems. Art. 38 of the DSA establishes an obligation according to which "providers VLOPs and of VLOSEs that use recommender systems shall provide at least one option for each of their recommender systems which is not based on profiling". Given the definition of profiling according to Art. 4 (4) GDPR, this in essence calls for the availability of a curation option that does not rely on the automated processing of personal data aimed at evaluating certain personal aspects relating to a natural person. While this represents an intuitive approach that strengthens the agency of users, a recent study on Facebook and Instagram usage during the 2020 US election did not find significant effects on the levels of polarisation or political knowledge when users' feed algorithms were switched from the default algorithmic curation mechanism to reverse-chronologically-ordered feeds.[118]

[114] Cf. Thorburn, L., Bengani, P., & Stray, J. (2022). What does it mean to give someone what they want? the nature of preferences in recommender systems. *Understanding Recommenders*. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157; Bundtzen, S. (2022). News Feed, Reels & »Für Dich«: Wie algorithmische Ranking-Praktiken unser Online-Umfeld beeinflussen und schützen könnten. https://www.isdglobal.org/isd-publications/news-feed-reels-fur-dich-wie-algorithmische-ranking-praktiken-unser-online-umfeld-beeinflussen-und-schutzen-konnten/

[115] In this sense, the pull-based mechanisms of search engines allow for such increased user agency over push-based recommender systems. Nonetheless, some of the measures to further increase user control could also be applied to search engines.

[116] Fukuyama, F., Richman, B., Goel, A., Schaake, M., Katz, R., & Melamed, D. (2020). Report of the working group on platform scale.

[117] Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K., Lewandowsky, S., Hertwig, R., ... & Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 1-9.

[118] Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science*, *381*(6656), 398-404. https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale

There are in general open research question about whether and under which conditions interventions to promote user control can indeed present effective mitigation measures and whether they can ultimately lead to beneficial outcomes. This is because giving users more control over their information space could even reinforce some of the concerns associated with automated information processing.[119] For example, users may use filter mechanisms to cut out information or content that is not aligned with their interests and preferences, thus possibly further exacerbating polarisation and radicalisation tendencies by lowering the exposure to diverse information and opinions. Furthermore, there is the open question of how such tools must be designed to be adopted and continually used by users. If individual-level interventions would rely on the integration of third-party tools this could also present opportunities for malign actors to target and exploit such direct user access and the ability to shape users' individual information space. Even without malign intent, the integration of third-party tools could present risks to privacy and data protection, as internal user information, possibly including sensitive personal and behavioural data, would need to be shared between the platform and external tool providers (see also Section 4.3). Hence, the integration of decentralised third-party tools could contribute to systemic risks on its own.

### *Explicit preference elicitation by survey questions*

Inferring user preferences exclusively from their observed behaviour has several shortcomings as detailed above. Therefore, a natural extension to engagement-based ranking metrics is giving more weight to user's explicit feedback elicited through direct survey questions.[120] This explicit feedback can be collected both on an individual or on a collective basis. Questions to users may concern choices between different suggested options or be more open-ended to collect broader feedback.

The advantage of explicit preference elicitation is that users can more deliberately reflect on a choice they face, the alternative options, and the implications involved. Hence, it can be expected that users will be more likely to make decisions based on conscious, rational thinking, and controlled mental processes rather than based on intuitive and affective reactions.[121] In consequence, user surveys can incorporate users' more long-term considerations and preferences as well as reflections on their past behaviour and experiences. Surveys can be conducted based on questionnaires that include multiple questions and cover broader areas, but the elicitation of explicit user feedback may also be integrated into the usual flow of user actions and the user interface. For example, users may be asked to reflect on the recent purchase or use of a product that they just made or to provide feedback about their perceived well-being in the context of a recent platform interaction (e.g., "Was this post worth your

[119] Bundtzen, S. (2022). News Feed, Reels & »Für Dich«: Wie algorithmische Ranking-Praktiken unser Online-Umfeld beeinflussen und schützen könnten. https://www.isdglobal.org/isd-publications/news-feed-reels-fur-dich-wie-algorithmische-ranking-praktiken-unser-online-umfeld-beeinflussen-und-schutzen-konnten/

[120] Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, *3*(4), 443-463; Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W. P., & Jonker, C. M. (2012). Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction*, *22*, 357-397.

[121] Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

time"?).[122] In addition, questions may concern users' retrospective evaluation of their own actions and associated experiences.

The main downside of the collection of explicit feedback is that it requires attention and effort from users. Hence, survey questions may be perceived as annoying interruptions of the actual service usage and answering them can be a time-consuming activity without an immediate benefit for the users. Moreover, collecting user feedback explicitly may be perceived as more intrusive than collecting implicit feedback, of which users are often not even aware. In consequence, this can raise privacy concerns or lead to user reactance in addition to the usual limitations of survey methods. Similar to user control, it is also not clear, whether explicit preference elicitation is generally effective at promoting exposure to more diverse information and at mitigating echo chamber effects, or if explicit feedback may even reinforce some of the concerns regarding automated information processing.

Explicit preference elicitation and integration of survey questions have a long tradition in the design of recommender systems.[123] With regard to concerns about automated information processing, there are indications that some providers of VLOPs and VLOSEs have augmented the role of user surveys and given them more weight for the recommendation and algorithmic curation of information.[124] This points to the potential of this approach as a risk mitigation mechanism, but as for the other approaches, there are many open questions about the requirements, context-specificity, and implications of such an approach that require more cross-cutting analysis and empirical evaluation in the actual application settings.

### *Bridging algorithms and diversity-enhancing techniques*

Bridging systems and bridging-based ranking have been proposed as possible mitigation mechanisms to reduce division and polarisation on online platforms, especially social media services.[125] The key idea is to "[reward] content that led to positive interactions across diverse audiences, including around divisive topics"[126] with the goal to "increase […] mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation."[127] According to this proposed

[122] Gupta, A. (2021). Incorporating More Feedback Into News Feed Ranking. https://about.fb.com/news/2021/04/incorporating-more-feedback-into-news-feed-ranking/; Sethuraman, R., Vallmitjana, J., Levin, J. (2019). Using Surveys to Make News Feed More Personal. https://about.fb.com/news/2019/05/more-personalized-experiences/ Thorburn, L., Bengani, P., & Stray, J. (2022). What does it mean to give someone what they want? the nature of preferences in recommender systems. *Understanding Recommenders*. https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157

[123] See, e.g., Bennett, J., & Lanning, S. (2007). The Netflix prize. *Proceedings of KDD Cup and Workshop 2007*; Liu, N. N., Xiang, E. W., Zhao, M., & Yang, Q. (2010). Unifying explicit and implicit feedback for collaborative filtering. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1445-1448).

[124] Stepanov, A., Gupta A. (2021). Reducing Political Content in News Feed. https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/; Stray, J. (2020). Aligning AI optimization to community well-being. *International Journal of Community Well-Being*, *3*(4), 443-463.

[125] Ovadya, A. (2022). Bridging-Based Ranking. How Platform Recommendation Systems Might Reduce Division and Strengthen Democracy. https://www.belfercenter.org/sites/default/files/files/publication/TAPP-Aviv_BridgingBasedRanking_FINAL_220518_0.pdf; Ovadya, A., & Thorburn, L. (2023.) Bridging Systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance. https://knightcolumbia.org/content/bridging-systems; Bundtzen, S. (2022). News Feed, Reels & »Für Dich«: Wie algorithmische Ranking-Praktiken unser Online-Umfeld beeinflussen und schützen könnten. https://www.isdglobal.org/isd-publications/news-feed-reels-fur-dich-wie-algorithmische-ranking-praktiken-unser-online-umfeld-beeinflussen-und-schutzen-konnten/

[126] Ovadya, A. (2022).

[127] Ovadya, A., & Thorburn (2023)

approach, engagement-based ranking metrics should be complemented (or even replaced) by bridging metrics that are aimed at promoting exposure to diverse opinions and reducing polarisation and divisive interactions. This is not supposed to eliminate conflict or to create homogeneity but to allow for productive conflict, deliberation, or cooperation.

Metrics for bridging-based ranking could build on survey questions that attempt to measure the extent to which recommended content is perceived as 'bridging' by users. For example, questions measuring partisan animosity or affective polarisation could be used to assess users' perception of recommended content. Given some explicit user feedback, this information could then be leveraged to predict users' perceptions of other content without the need to gather feedback on each individual piece of information. Furthermore, user behaviour can serve as implicit feedback and allow for the classification of 'bridging motifs', which capture behavioural patterns conducive to bridging outcomes. To this end, Ovadya and Thornburn discuss diverse approval (i.e., an item is endorsed by people from multiple diverse viewpoints), response bimodality (i.e., the distribution of ratings or reactions to an item is not polarised), or exposure diversity (i.e., people attend to items from diverse sources, particularly from sources they don't normally see) as potential motifs that could serve as ranking signals for bridging systems.[128]

While there are instances where providers of VLOPs have integrated elements of bridging systems into recommender systems,[129] the effectiveness of bridging systems in achieving the intended goals as well as their wider implications for the user behaviour and collective outcomes is largely unknown. Despite their intuitive appeal, there are also several major challenges and possible risks associated with their implementation. Most notably, there needs to be some judgment on what outcome benchmarks and associated metrics should be deemed as indicators of 'beneficial' outcomes. Already the presumption that services should promote interactions between diverse audiences is not shared by everyone. Operationalising the bridging goal and breaking it down into actual metrics will be even more contentious. In consequence, this calls for institutions and processes that could guide the possible development and evaluation of bridging-based elements.[130]

Furthermore, it is not clear how users will respond and react to bridging-based systems. In particular, there is the risk that users perceive such recommendations as paternalistic or simply as not relevant to their interests. In consequence, this could lead to lower service usage or user churn, which will then also affect the incentives of the providers of such services to implement these systems. As a further implementation challenge, it is also not obvious, which types of risks and which specific services would be particularly suited to adopt elements of bridging-based systems. For example, whereas social media services commonly used for political debates and opinion formation may be seen as primary

[128] Ovadya, A., & Thorburn, L. (2023.) Bridging Systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance. https://knightcolumbia.org/content/bridging-systems

[129] See, e.g., Community Notes on X, which displays a user-generated note on any post if enough contributors from different points of view rate that note as helpful. See https://help.x.com/en/using-x/community-notes

[130] Cf. Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., Ashar, A., ... & Vasan, N. (2024). Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, *2*(3), 1-57.

candidates for adopting bridging systems, it is far less clear whether bridging approaches should also be considered for social media services focused on employment or the sharing of everyday activities.

As the DSA itself does not further distinguish between the large diversity of VLOPs and VLOSEs, additional assessments seem to be warranted to develop generalizable criteria for when bridging-based rankings and related approaches could and should be considered as potential measures to mitigate systemic risks. In this context, our proposed differentiation of distinct service types could be instructive (see Section 2.2). Even for the same type of systemic risk, a differentiated approach may be advisable. For example, for risks to electoral processes, some periods in the election cycle may call for more ‘bridging interventions’ than others, which implies that the involved trade-offs could be balanced differently for different contexts and points in time. Overall, this further reiterates the need for i) more empirical evaluation of what mitigation measures and corresponding designs of technical interventions can actually have a positive impact and ii) a coordinated approach to assess and guide the evaluations and the implementation of mitigation measures across individual VLOPs and VLOSEs.

# 5. RECOMMENDATIONS

The overarching suggestion we make based on this analysis is that there is a need for meta-analysis across the individual services' risk assessments, in addition to the compliance focused evaluation of each of them. The following more specific recommendations are ones that we think the European Commission as regulator of VLOPs and VLOSEs and the Digital Services Board (DSB), with its wide convening power and advising, should consider. In these recommendations there is also a call to academics and civil society groups to engage in meta-analysis of the risk assessment reports once they are public and to make strategic use of the access provided to vetted researchers through Article 40 of the DSA. In this vein, the meta-analysis of risks is intended to be performed as a cooperative endeavour that enables iterative learning over subsequent risk assessment cycles and that will require the participation of the providers of VLOPs and VLOSEs, regulators, researchers, and various stakeholders.

1. An inclusive, cooperative process should take place, possibly led by the Digital Services Board and the Commission to set priorities among the risk areas for meta-analysis, taxonomy of harms, and strategies for consistent use of information-gathering tools across services and over time, including setting and periodically reviewing standards for the relevant metrics and data.

2. For each specific category of systemic risk set out in Article 34 of the DSA, meta-analysis across risk assessments should aim to harmonize the definitions of core concepts relevant to the risk area and the negative effects to be prevented, understandings of norms and policy goals, and data gathering and reporting.

3. The following specific areas for meta-analysis across services by independent researchers and Digital Services Coordinators through coordinated use of data access and publicly reported information including in risk assessment reports and the transparency database:
   a. Advertising business models: effects of targeting; effectiveness of ad libraries;
   b. Temporality features: possible correlation with malign use
   c. Use of automated cross-posting tools;
   d. Very large influencers and related mitigations;
   e. Effectiveness of control and incentive mechanisms, and combinations thereof, on specific sources of risk;
   f. Recommender systems: transparency and effects of ranking signals and algorithmic curation decisions on user behaviour and collective outcomes;
   g. Inauthentic use and generative AI;
   h. Data sharing, data agglomeration, and common critical technical vulnerabilities;
   i. The roles of users, third parties and common resources or assets in content moderation.

4. As exemplified by our analysis, the meta-analysis of potential mitigation strategies to identify best practices (possibly within the categories identified in this report), which could make use of existing data but also may require experimentation, should include:
    a. An evaluation of the effectiveness of extensions to engagement-based recommender systems and algorithmic curation systems, specifically user control, explicit preference elicitation, and bridging-based algorithms.;
    b. An examination of the benefits and vulnerabilities from decentralisation especially of content moderation and governance of user behaviour, and the outcome of various balances and mixes between decentralised and centralised governance mechanisms;
    c. Particular attention to aversion risk and user flight or changes in user of servicer in response to mitigations in the wider interconnected digital services landscape.

5. Building on existing cooperation mechanisms for particular types of harm (CSAM, terrorist content, disinformation, hate speech), identify opportunities where additional cooperation for risk mitigation between service providers and stakeholders is needed. For instance, through rapid response mechanisms, codes of conduct, incident databases and coordinated intelligence gathering and mitigation approaches to safeguard against influential malign users.

Centre on Regulation in Europe

Avenue Louise 475 (box 10)
1050 Brussels, Belgium
+32 2 230 83 60
info@cerre.eu
www.cerre.eu

Centre on Regulation in Europe (CERRE)
CERRE Think Tank