Initial Analysis on the First Round of Risk Assessments Reports under the EU Digital Services Act

An initiative of the DSA Civil Society Coordination Group

March 2025

Introduction

In this brief, the DSA Civil Society Coordination Group, in cooperation with the Recommender Systems Taskforce[1] and People vs Big Tech[2], provide an initial, high-level analysis of the first round of Risk Assessment Reports (RA Reports), published under Art. 42 of the Digital Services Act (DSA).

The risk assessments and subsequent mitigation measures form a crucial part of the due diligence framework for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) under the DSA. Since the entry into force of the DSA, it has been noted that the nature of this assessment and reporting exercise is iterative and that it is a 'learning exercise'[3]. The intent, therefore, is to continuously improve upon each risk assessment and mitigation process, as well as on the way these are reported on, in order to achieve meaningful transparency.

The purpose of this initial feedback is to identify key trends, useful practices and gaps in this first iteration, so that future iterations of these reports advance the public interest, particularly at a time of diminishing trust amongst users related to the risks that VLOPs and VLOSEs create at the personal and societal level[4]. Companies seeking to demonstrate effective compliance with the DSA and to foster trust with their users should take this feedback into consideration.

This brief will focus on four key aspects: ***useful practices from the RA Reports; the importance of focusing on platform design when assessing risk; why trust can only be fostered through transparency; and the need for meaningful stakeholder engagement***. Due to the prevalence of their services in the EU, as some of the largest of the 25 very large online platforms and search engines, the focus is on the RA Reports of the following companies: Google (Search and YouTube); Meta (Facebook and Instagram); TikTok and X.

Civil society and researchers remain willing to provide more detailed insights, analysis and constructive feedback on the risk assessment methodology and development of mitigation measures in a manner that is meaningful and equitable. In this regard, it is important that in the future, VLOPs and VLOSEs proactively reach out to organisations and researchers in the EU as part of these processes.

---

[1] https://civitates-eu.org/grantee-spotlight-the-panoptykon-foundation-recommender-systems-task-force/
[2] https://peoplevsbig.tech/
[3] For example, Hendrix J. & Miller G. (2024) Assessing Systemic Risk Under the Digital Services Act, Tech Policy Press; Broughton Micova S. & Calef A. (2023) Elements for Effective Risk Assessment Under the DSA, CERRE but also, for instance, Renate Nikolay (Deputy Director of DG CNECT) during the event 'The Perfect Storm'.
[4]For example, IAPP (2024) Users' trust in social media companies declining referencing the Thales Digital Trust Index ; Bursztyn L., Handel B.R., Jimenez R. & Roth C. (2023) When Product Markets Become Collective Traps: The Case of Social Media, National Bureau of Economic Research

1. Useful practices and considerations for future iterations

As expected, the first round of RA Reports revealed that companies have chosen different approaches when it comes to identifying and assessing risks, and choosing how to mitigate them. There are also notable differences in the formats chosen to convey this information. While the reports across platforms and services will inevitably differ to some extent, we have identified practices employed by certain providers in this first iteration that other providers should emulate, as a minimum, where appropriate in the future. We also point to certain practices that do not feature in the RA Reports analysed, but should be considered for future iterations. In this section, we have chosen to focus on risks to the mental health of children, media pluralism and online GBV as examples, but our observations and recommendations are not limited to these specific risks alone.

1.1 Assessing the risk to the mental health of children, in particular focused on addictive design

Some independent research shows that addictive features of platforms' design can lead to problematic (over-)use of online services by minors[5], which in turn has been linked to negative mental health outcomes[6]. The evidence behind the addictive design of platforms has led some governments to consider banning social media platforms for people under 16 years old altogether[7], while the DSA explicitly requires VLOPs and VLOSEs to address risks which may cause addictive behaviour, especially for minors[8].

In Google's RA Report, it is explicitly mentioned that the features and design of YouTube may stimulate behavioural addictions in children[9]. Accordingly, YouTube implemented mitigation measures such as 'Take a Break' notifications and bedtime reminders, as well as disabling "autoplay" automatically for supervised accounts and in the YouTube Kids App in order to reduce that risk.

---

[5]Chapman P. (2024) Advancing Platform Accountability: The Promise and Perils of DSA Risk Assessments, Tech Policy Press; Langvadrt K. (2019) Regulating Habit-Forming Technology, Fordham Law Review; Lukoff K. et al. (2021) How the Design of YouTube Influences User Sense of Agency, Association for Computing Machinery; Bernstein G. (2023) Unwired – Gaining Control over Addictive Technologies, Cambridge University Press; Zhang et al. (2021) Ephemerality in Social Media: Unpacking the Personal and Social Characteristics of Time Limit Users on WeChat Moments, Frontiers in Psychology ;

[6] Caplains, S.E. (2010) Theory and measurement of generalized problematic Internet use: A two-step approach, Computers in Human Behavior; Casale, S. & Banshee, V. (2020) Narcissism and problematic social media use: A systematic literature review, Addictive Behaviours Reports; Bányai, F. et al (2017) Problematic Social Media Use: Results from a Large-Scale Nationally Representative Adolescent Sample, PLoS ONE; Brautsch L.A.S. et al. (2023) Digital media use and sleep in late adolescence and young adulthood: A systematic review, Sleep Medicine Reviews

[7]Ritchie, H. (2024) Australia approves social media ban on under-16s, BBC

[8] DSA Recital 81

[9] Google (2024) p.106 "YouTube considered numerous risks particular to children [...] the risk that YouTube stimulates behavioural addictions in children"

Given that the key features and design elements[10] that have been linked to addictive behaviour are not unique to YouTube, but are present in other platforms, we believe that in accordance with DSA Art.34 (1)(d), all social media platforms should include a clear, explicit and specific assessment of the risks linked to addictive design. They must also include accompanying mitigation measures, which should not be limited to those that YouTube has implemented. It is worth noting that even in the case of Google, neither the RA Report, nor YouTube's audit report mention how the efficacy of the aforementioned mitigation measures was evaluated, which means that it cannot be determined whether YouTube have successfully mitigated the risk.

Though it is encouraging that at least one platform has explicitly considered the addictive effect that its design and features may have on children, it should be noted that all other RA Reports that we analysed failed to sufficiently assess or mitigate this risk. For example, Facebook's RA Report mentions what they refer to as "problematic use" of their services in a way that seems designed to downplay the importance and severity of that risk[11]. Instagram's RA Report mentions that "[u]sers are encouraged to access time management tools and resources to have better control and feel more agency over their experience"[12] as a mitigation measure, without specifically citing what risk this is addressing. To understand what time management tools Instagram has introduced, one must go to a different section and follow an external link[13], then search and find a section titled 'support tools' to find the right page[14] – only then are specific mitigation measures such as 'setting a daily time limit' described, though their efficacy is not elaborated upon as those are public resources and were not designed for the purpose of complying with the DSA.

In a similar fashion, TikTok implements screen time management tools as mitigation measures in relation to the risks associated with the online protection of minors[15], without directly acknowledging how the design of its services itself may contribute to the addictive behaviour that time management tools are supposed to curb. Most worryingly, it is alleged that its own internal documents show that the tool had little impact on the time spent by teens on the service[16].

It is concerning that such a prominent risk, that is also clearly detailed in the DSA, features so obscurely in the RA Reports of providers whose services are used by millions of minors in the EU. Meta justifies this lack of specificity by explaining that mental and physical risks are assessed in a

[10] Langvadrt K. (2019) Regulating Habit-Forming Technology, Fordham Law Review; Lukoff K. et al. (2021) How the Design of YouTube Influences User Sense of Agency, Association for Computing Machinery; Bernstein G. (2023) Unwired – Gaining Control over Addictive Technologies, Cambridge University Press; Zhang et al. (2021) Ephemerality in Social Media: Unpacking the Personal and Social Characteristics of Time Limit Users on WeChat Moments, Frontiers in Psychology

[11] In fact, the term "problematic use" is only used once in the 95-page document, not as a risk, but rather as context for a mitigation measure Facebook (2024) p. 50.

[12] Facebook (2024) p. 50; Instagram (2024) p. 49 "Meta encourages users to maintain well-being tactics on Instagram through various time management tools and topic switching measures."

[13] Instagram (2024) p.10 https://about.instagram.com/safety/account-safety#support-tools

[14] https://help.instagram.com/195902884574087?helpref=uf_permalink

[15] TikTok (2023) p. 21

[16] Allyn B. & Kerr D. (2024) TikTok executives know about app's effect on teens, lawsuit documents allege, NPR

crosscutting way, contributing to all risk areas – the result of this approach, however, is that specific and important mental health risks such as the problematic over-use of a service are almost completely overlooked.

Beyond explicitly identifying and focusing on the risk of problematic over-use of a service, any mitigation measure aimed at countering this risk should also be targeted at mitigating the risk of developing an addiction – as in the case of YouTube – and not merely the risks stemming from the addictive use of the service. To illustrate the latter, X's RA Report mentions that "there is a risk that heavy usage of social media may lead to increased risk for depression, anxiety, social isolation, self-harm, and suicidal thoughts"[17]. The mitigation measure that is then cited in relation to this risk is that "X has developed a policy prohibiting users from promoting or encouraging suicide or self-harm"[18]. This mitigation measure does little, if anything, to mitigate the stated risk of "heavy usage".

In conclusion, if providers care about the physical and mental health of their users, particularly those who are considered minors, and in order to comply with article 34 (1)(d) and recital 81 of the DSA, they must do more. They should clearly and explicitly assess the risk of users developing addictive behaviour, apply reasonable, proportionate and effective mitigation measures and explicitly report on those measures with data substantiating their claimed effectiveness.

1.2 Consideration of media pluralism as integral to a healthy information environment

Reference to media pluralism in respect to promoting free and open access to information was a welcome finding within a handful of RA reports[19] and should feature prominently in future iterations as well. Within the framework of fundamental rights, media pluralism falls under the right to free expression and is a prerequisite to human development, dignity, the search for truth and the driver to the enjoyment of all other rights.[20]

For example, in its RA Report, X leads its section on Fundamental Rights by citing inherent risks to freedom of expression and information, namely by citing the different ways that information from transparent and pluralistic sources can be inhibited on the platform.[21]

---

[17] X (2023) p. 68
[18] X (2023) p. 69
[19] In addition to references by X and Google Search noted below, Bing not only cites support for media pluralism as a foremost commitment throughout its RA report, but also highlights the provision of a "variety of high authority news resources and [establishment of] safeguards to avoid inadvertently presenting users with low authority content" as a mitigation measure. Bing 2024 p. 135.
[20] For example, Centre for Law and Democracy & IMS (2014). Freedom of Expression Briefing Note Series ; UNESCO. (2023). World Press Freedom Day 2023 Draft Concept Note: Shaping a Future of Rights, Freedom of expression as a driver for all other human rights.
[21] X (2023) p. 36

Google Search references respect for media pluralism as a mitigation measure in the context of investing in information quality.[22] Within the same section, elevating authoritative information and combating mis- and disinformation are also cited as critical to addressing systemic risks on the search engine. While these insights provided in Google Search's RA Report would benefit from further detail, such considerations should be seen across all platforms in future iterations.

A notable omission across most of the RA Reports reviewed pertains to a missing analysis on the promotion of independent journalism as a mitigation measure to uphold information integrity. As it is not only the minimisation of disinformation, hate speech and other concerning communication, but also the maximisation of accurate and reliable information that contributes to a well-functioning democratic society, a plurality of journalistic sources is a mitigation measure that must not remain overlooked. Public interest media is generally grounded in ethics and appreciates the importance of presenting content that reflects how their audiences can practise democratic citizenry.[23] A variety of vetted journalistic sources should therefore be considered integral to contributing to a healthy information environment and we hope to see this reflected in future RA Reports.

1.3 Mitigation measures in relation to Online Gender-Based Violence (OGBV)

In the RA Reports by X and Meta, specific risks are identified in relation to OGBV and accompanying mitigation measures are implemented. The two companies, however, have since changed some of their policies and rolled back some of those mitigation measures by providing reasoning for these rollbacks that have been questioned by civil society. We believe that the risk assessment and mitigation practices featured in the RA Reports by X and Meta should not have been discontinued and must be reinstated, at a minimum, in future iterations.

X's risk assessment outlines controls to deal with the risk of gender-based violence which includes 'Safety features: Features such as block/mute, account filters, and controlling replies can protect users from gender-based violence (GBV)'[24]. In 2024, however, X introduced changes to its blocking feature, which now allow blocked accounts to see a user's public posts, but not interact with them, in a move that was widely criticised by civil society as increasing the risk to women and girls from abusers[25].

Meta's RA Reports for both Instagram and Facebook cite that '[w]e have worked to strengthen our relationships with the LGBTQIA+ community by increasing engagements with groups and representatives across the world and in the EU, on the impact of Meta's content policies on users,

[22] Google (2024) p. 71-72
[23] Ag, M., Refsing, N.S, & Lehmann-Jacobsen, E. (2023). Public interest infrastructure: Digital alternatives in our data-driven world and journalism's role getting there International Media Support
[24] X (2023) p. 70-71
[25] Landi M. (2024) X accused of 'lack of care' for women and girls over blocking feature change, The Independent

particularly regarding hate speech, bullying, and harassment'[26]. However, it is questionable how seriously Meta took the recommendations these groups may have given them, especially in light of recent developments. In January 2025, it was reported that Meta enacted changes to its Hateful Conduct policy, which have been condemned by LGBT+ groups around the world[27]. Similarly, Meta says that its commitment to women's safety is longstanding and that it has developed strong policies to help protect women from online abuse. In the same hateful conduct policy updates, however, Meta includes specific carve outs to enable greater levels of hateful speech to be targeted at individuals on the basis of sex and gender[28].

The initial assessments conducted and mitigation measures implemented by the two providers in relation to OGBV were a step in the right direction and should (at the very least) become once again part of the way that X and Meta assess and mitigate risk on their platforms. If measures to effectively combat OGBV are not reinstated, then the two providers in question will have to substantiate specifically what assessments were conducted in order for X and Meta to make the determination that these measures were no longer required. We believe that the European Commission should probe further in relation to the decisions X and Meta have made in this regard[29].

1.4 Assessing the risk to Fundamental Rights

The trend of omitting or failing to mitigate against specific systemic risks, as illustrated in the sections above, extends beyond the examples provided. With this in mind, it is pertinent to remind VLOPs and VLOSEs that they are required to assess and mitigate a breadth of potential risks, and not ignore or fail to report on those they deem irrelevant. Article 34.1(b) of the DSA requires an assessment of "*any actual or foreseeable negative effects for the exercise of fundamental rights"*. This means that all VLOPs and VLOSEs should assess the risk their products and services pose to *any and all* of the rights affirmed by the EU Charter of Fundamental Rights, though particular attention should be paid to the rights to privacy, freedom of expression and information, and to non-discrimination, among others. Though it has been recognised that achieving this effectively is a challenge, civil society has offered guidance on how VLOPs and VLOSEs can carry out a fundamental rights impact assessment which build on existing human rights impact assessment methodologies[30].

Unfortunately, however, and as will be referenced throughout this brief, our initial analysis of the published RA Reports is that several providers have either failed to conduct this broader

[26] Instagram (2024) p. 23 ; Facebook (2024) p. 23
[27] Torek B. (2025) Meta's New Policies: How They Endanger LGBTQ+ Communities and Our Tips for Staying Safe Online, Human Rights Campaign; Blake S. (2025) Stonewall responds to Meta's new policy changes, Stonewall
[28] Meta Transparency Center Hateful Conduct
[29] In addition to the RfIs sent to X on Oct. 12th 2023 and 8 VLOPEs on Mar. 14 2024.
[30] ECNL & AccessNow (2023) Towards Meaningful Fundamental Rights Impact Assessment Under the DSA

assessment or have chosen not to specify how this assessment was conducted, and their resulting conclusions. In future iterations of the RA reports, it would be beneficial for platforms to provide more insights on if and how they have conducted fundamental rights impact assessments, or assessments utilising a similar methodology, as part of their risk assessment and mitigation efforts.

1.5 Readable and digestible format

To serve as a meaningful tool for transparency, the RA Reports also need to be published in a format that is digestible and – eventually – comparable. Though there will always be differences in the reporting formats, reflecting the different nature and features of each platform service, we have identified certain practices that make it easier for independent experts, the public and the regulators to understand, evaluate and compare the reports.

1. **Length and specificity** – Given that the reports serve the purpose of demonstrating compliance with the DSA, it is understandable that they need to be thorough. At the same time, we have noted instances where the length of certain reports could be reduced, if the content was more specific to the DSA. For instance, the first half of Google Search's RA Report describes many general Google social responsibility measures, without being clear on how these are relevant to the DSA[31]. It appears likely that for the first iterations of these reports companies have relied heavily on existing material, instead of producing new, tailored analyses of their services under the DSA. Going forward, the information provided in the RA Reports should be specific to the DSA, which should also limit the length and make the reports more digestible.
2. **Visualisation and structure** – The way that reports are structured and visualised also plays a crucial role in how a reader is able to navigate them. TikTok's RA Report, which uses similarly structured tables for every risk and related mitigation measures makes it much more digestible and we encourage similar visual approaches.
3. **Hyperlinks and references** – The use of hyperlinks is encouraged for referencing relevant research, such as in the case of the RA Reports by X[32] and TikTok[33]. However, specificity is crucial in terms of how a reference serves to substantiate a particular statement or claim. A generic sentence, followed by a hyperlink, which implies that more information can be found in the page that is linked, adds little to the reports. Rather, it makes it difficult for the reader to follow the precise reasoning, explanation or evidence that is being showcased. This is exemplified by Meta's approach to addressing the

[31] Google Search (2024)
[32] For example, X(2023) p. 55-56. While X's report is strikingly limited in multiple ways, it is one of the few that provides some evidence for empirical grounding. It cites research to support its decisions, though the credibility of the referenced studies remains debatable. For instance, while X's heavy dependence on Community Notes as its primary tool to combat misinformation is questionable, the report at least attempts to justify this approach with a rationale.
[33] For example, TikTok (2023) p. 11

'problematic use' of its services in section 1.1 of this brief. Along the same vein, we often saw platforms link to publicly available transparency reports to reference DSA-compliant data, but without specifying where in the report the relevant information could be found, making it very difficult for readers to locate the referred data.

4. **Rubrics and metrics** – In their RA Report, Meta acknowledges the importance of metrics[34] and the rubric for risk-levels they include details how many people were exposed to a certain risk[35]. Like Meta, all platforms must fully acknowledge the importance of metrics. However, all platforms including Meta, should also provide those metrics publicly. A common issue across platforms was the lack of insightful metrics around exposure to harmful content and quantifiable mitigation metrics mapped to key risk dimensions like Scale, Cause, and Nature. Rubrics and metrics should be presented in clear side-by-side visualizations to promote better comparative readability.
5. **Machine readable data** – Though this is not required by the DSA and therefore no provider has done it, having machine-readable documents to go along with the RA Reports would help extract and analyse relevant data. The transparency reports for the Code of Practice on Disinformation have such machine-readable documents (JSON and CSV), which include all the data points.

2. Why platform design must not be overlooked

A key trend across platforms is that the RA Reports disproportionately focus on content and user-generated risks, while overlooking design-related risks[36]. The risks posed by the design of services are occasionally acknowledged, but rarely referred to directly, though it is clear that providers are aware that design is a crucial element in analysing systemic risks, as is also highlighted in the DSA.[37]

References to design – and in particular recommender systems – can be found in the reports, although they are comparatively underrepresented in relation to user-generated risks. Some notable, albeit few, examples of design-related risks are included here. Bing acknowledges that the design of its platform and in particular their recommender systems can horizontally contribute to a variety of systemic risks[38]. Google acknowledges that tailoring recommender systems to recommend content based on the quality of the content, rather than solely on engagement, can

[34] For example, Instagram (2024) p. 31, 46
[35] For example, Instagram (2024) p. 88-93
[36] This has been noted by several commentators, for instance during a recent event 'Decoding DSA Risk Assessments and Audits at LSE' with Agne Kaarlep and Martin Husovec, or by Chapman P. (2025) Advancing Platform Accountability: The Promise and Perils of DSA Risk Assessments, Tech Policy Press.
[37] DSA recitals 79, 81, 83, 84, 87 and Articles 34(1) and 34(2)(a)
[38] Bing (2023) p.25 "Bing invests significant time and resources into ensuring its crawlers and algorithms prioritize high quality content to avoid inadvertently returning low quality or harmful content to users."

be an effective mitigation measure in and of itself[39]. Meta acknowledges that limiting the role of shares and comments in the distribution of sensitive topics can be an effective mitigation measure[40]. To the best of our knowledge, only TikTok fails to acknowledge this explicitly, instead saying that their "content is served based on interests and user engagement so entertainment is always personal"[41].

Taken collectively these approaches show a general reluctance to recognise the central role that design plays in the creation of risks, referring to it only tangentially, or indirectly by pointing to mitigation measures. At the same time, mounting evidence from independent, authoritative research continues to shed light on how different design features may contribute to different kinds of risks[42]. This is crucial when it comes to the physical and mental health of children, as already outlined in section 1.1, but has cross-cutting implications at the individual and societal level for all systemic risk areas.

In particular, recommender systems that rank content based on engagement have been shown to amplify problematic content in ways that contribute to various systemic risks[43]. To illustrate this by referring to a specific platform, independent research has documented systemic mental health risks in relation to "rabbit holes" associated with TikTok's 'For You Feed'[44], which TikTok fails to address in its RA Report, other than by claiming that it "employs mitigation measures to diversify content so that Younger Users are not exposed to repetitive content, which is especially important if they are exploring content related to more complex themes, but which is not in violation of TikTok's terms or Community Guidelines"[45]. In the same report, mental health risks are only discussed in connection with suicide challenges and hoaxes, much like most RA Reports analysed, which focus on content risks at the expense of addressing design risks. Providers therefore appear to knowingly omit systemic risks, which could contribute to significant offline

[39] Google (2024) p.25 "Using recommender systems to order the presentation of content, including by elevating high-quality and trustworthy content, is a more proportionate approach to addressing harmful content risk than removing content altogether"

[40] Meta (2024) p.29 "We routinely evaluate whether the signals we use to enable users to get relevant content could lead to exposure of problematic content. We reduce this risk by limiting the role of shares and comments in the distribution of sensitive topics"

[41] TikTok (2023) p.3

[42] National Academies of Sciences, Engineering, and Medicine (2024) Social Media and Adolescent Health Office of the Surgeon General (2023) Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory [Internet] ; eSafety Commissioner (2019) Safety by Design Overview ; Cunningham T. et al. (2024) What We Know About Using Non-Engagement Signals in Content Ranking ; KGI Expert Report (2025) Better Feeds: Algorithms That Put People First

[43] Integrity Institute (2024) On RIsk Assessment and Mitigation for Algorithmic Systems ; KGI Expert Report (2025) Better Feeds: Algorithms That Put People First ; Bavel et al. (2021) How Social Media Shapes Polarization, Trends in cognitive sciences ; Brailovskaia et al. (2022) Experimental Longitudinal Evidence for Causal Role of Social Media Use and Physical Activity in COVID-19 Burden and Mental Health, Journal of public health ; Park et al. (2020) Global Mistrust in News: The Impact of Social Media on Trust, International Journal on Media Management; United Nations (2024) Independent Investigative Mechanism for Myanmar, "Anti-Rohingya Hate Speech On Facebook" ; Cunningham T. et al. (2024) What We Know About Using Non-Engagement Signals in Content Ranking

[44] Hobbs, T.D., Barry B. and Koh Y. (2021) 'The Corpse Bride Diet': How TikTok Inundates Teens With Eating-Disorder Videos, The Wall Street Journal ; CCDH (2025) Deadly by Design ; Eko (2023) Suicide, Incels, and Drugs: How TikTok's deadly algorithm harms kids

[45] TikTok (2023) p.23

harm in people with pre-existing mental health issues.

In the same vein, several reports focus on “bad actors’ behaviour” and combating the spread of violating content, but fail to address risks stemming from the hyper-personalisation of recommended content, which may not be dangerous (and therefore not eligible for moderation), but becomes harmful if consumed too much or by certain vulnerable individuals. Studies suggest that engagement-based recommender systems which rely on extensive collection and analysis of user data may over-expose users to large volumes of different kinds of legal (but harmful) content based on inferences about their individual sensitive features, fears or vulnerabilities[46]. This may trigger “unhealthy” engagement and push some users into harmful “doom-scrolling traps”, affecting their wellbeing or even exacerbating pre-existing mental health issues[47].

2.1 Risks for Minors

Minors are a particularly vulnerable group impacted by the design of platforms. For this reason, risks to the rights of the child are explicitly mentioned in article 34(1)(b) of the DSA along with mitigation measures such as targeted measures to protect the rights of the child foreseen by article 35(1)(j). The reports that we analysed fail to consistently consider the full range of risks minors face online and the need to adapt the design, features or functioning of online platform services and their algorithmic systems.

Firstly, several services rely on the assertion that their service is not aimed at minors nor predominantly used by them. On page 7 of its RA Report, TikTok asserts that “it is not specifically aimed at minors or predominantly used by them”, while many studies have determined that TikTok is one of the most used platforms by minors and its own internal documents show that TikTok considers users under 13 a “critical demographic”. In its audit implementation report, X declares that it “is not a service that is targeted at or predominantly used by minors, who represent a very small proportion of X account holders”[48] to dismiss the recommendations presented by the auditor. In any event, the fact, or otherwise, that the service is not predominantly used by minors does not relieve platforms of all responsibility.

In terms of the identification of risks, TikTok relied on the OECD 4Cs framework. However, TikTok failed to consider contract risks and cross-cutting risks, notably leaving privacy risks and risks of commercial exploitation un-identified and un-assessed. Overall, there is a broad failure to

[46]Integrity Institute (2024) Why Is Instagram Search More Harmful Than Google Search? ; CCDH (2025) Deadly by Design ; Amnesty International (2023) Driven into Darkness: How TikTok’s ‘For You’ Feed Encourages Self-Harm and Suicidal Ideation ; Hobbs, T.D., Barry B. and Koh Y. (2021) ‘The Corpse Bride Diet’: How TikTok Inundates Teens With Eating-Disorder Videos

[47] Amnesty International (2023) Driven into Darkness: How TikTok’s ‘For You’ Feed Encourages Self-Harm and Suicidal Ideation ; Panoptykon Foundation (2021) Algorithms of trauma: new case study shows that Facebook doesn’t give users real control over disturbing surveillance ads

[48] X(2023) p. 50

recognise and assess risks related to the design of services, in particular in relation to recommender systems. For instance, while internal TikTok documents allege the social media company is promoting addictive design, such risks are not addressed in its risk assessment[49].

Similarly, in terms of mitigation measures, the focus remains on content and content moderation, with reactive systems further placing the burden on users to identify and report on issues. In TikTok's risks mitigation table relating to its recommender system and its impact on children, the measures provided are about content moderation and transparency rather than the adoption of algorithmic systems. There is an additional lack of evidence relating to the effectiveness of the mitigation measures. While TikTok mentions the "Daily Screen Time" Management dashboard, internal documents showed that the tool had little impact on the time spent by teens on the service[50]. This further highlights the importance of companies providing the metrics associated with the effectiveness of the mitigation measures that have been implemented, as also mentioned in sections 1.4 and 3.

A recent study[51] further highlights how platform design — specifically YouTube's recommendation system — can actively contribute to systemic risks, directly pointing to key omissions in the platform's RA Report. By simulating a real-world scenario — a fictional 13-year-old in Ireland watching eating disorder-related content for the first time — the study exposes how YouTube's algorithm does not mitigate risk but instead amplifies it. Rather than steering users away from harmful content, the platform's design recommended more of it: one in three suggested videos contained harmful eating disorder content, while nearly three in four focused on eating disorders or weight loss.

Much like in the case of TikTok, YouTube's failure is not merely about gaps in content moderation; it underscores the deeper issue of design choices that prioritise engagement over user well-being. The study, however, also revealed that YouTube failed to remove, age-restrict, or label the vast majority of flagged harmful videos when accessed from an EU account—despite its own policies. Furthermore, YouTube's claimed risk mitigation measures, such as crisis resource panels, were found to be inconsistently applied, appearing in only two out of 27 EU countries. This left over 224 million users without access to critical support resources.

Such leaked documents and independent findings reinforce the urgent need to scrutinize how platforms design their systems, not just how they moderate content. In addition, without independent research and external accountability, platforms can present incomplete or misleading assessments of their own risk mitigation efforts—obscuring the fact that their core design choices may be driving harm. This study, like others before it, highlights that engagement-based recommender systems can systematically expose vulnerable users to content that exacerbates pre-existing issues, further demonstrating why platform design must be

[49]Allyn B. & Kerr D. (2024) TikTok executives know about app's effect on teens, lawsuit documents allege, NPR
[50] Allyn B. & Kerr D. (2024) TikTok executives know about app's effect on teens, lawsuit documents allege, NPR
[51] CCDH (2025) YouTube's EU Anorexia Algorithm: How YouTube recommends eating disorder videos to young girls in Europe

at the center of future RA Reports.

2.2 Civic Discourse and Electoral Processes

Unlike other systemic risks that platforms must independently identify, assess, and mitigate under the DSA, the European Commission has provided comprehensive, detailed guidelines outlining measures VLOPs and VLOSEs should take, in case they are reasonable, proportionate and effective in the given context, to safeguard electoral integrity[52]. Most of the RA Reports published in late 2024 cover periods preceding the publication of these guidelines, meaning that the full extent of use of the recommended measures remains to be seen[53]. Nonetheless, a recurring flaw in the risk assessments is the failure to critically examine how platform design and functioning influences civic discourse and electoral processes.

Platforms frequently frame risks as external threats posed by bad actors rather than systemic issues embedded in their recommendation algorithms, content moderation policies, and engagement-driven ranking mechanisms[54]. This framing often contradicts the evidence stemming from platforms's own research, as exemplified by a leaked internal report from 2020, which showed that 64% of joins to violent extremist groups on Facebook came from their recommendations[55]. Meta's reliance on its existing *Community Standards* to assess risks on Facebook and Instagram further exemplifies this problem. By anchoring risk identification in predefined policy violations rather than a fresh, systemic analysis, Meta avoids deeper scrutiny of how its algorithmic choices contribute to political polarisation, echo chambers, and the suppression of legitimate political speech. Meta has recently removed the contributions from shares and comments when ranking political content[56], which proves that they recognise how their algorithms can play a negative role, but fail to do so more comprehensively throughout their service.

More specifically, the design and functioning of a service can also play a crucial role when it comes to how political content is disseminated on a platform, with potential implications for systemic risks to electoral processes and civic discourse[57]. The European Commission's guidelines on the mitigation of systemic risks for electoral processes pursuant to Article 35(3) specifically call for platforms to "adapt their recommender systems to empower users and reduce the monetisation and virality of content that threatens the integrity of electoral processes" while also stating that "political advertising should be clearly labelled as such, in anticipation of

[52] Commission guidelines under the DSA for the mitigation of systemic risks online for elections
[53] EPD (2025) Civic Discourse and Electoral Processes in the Risk Assessment and Mitigation Measures Reports Under the DSA: An Analysis
[54] Stray J. Iyer R. & Puig Larrauri H. (2023) The Algorithmic Management of Polarization and Violence on Social Media, Knight First Amendment Institute at Columbia University
[55] Horwitz J. & Seetharaman D. (2020) Facebook Executives Shut Down Efforts to Make the Site Less Divisive, The Wall Street Journal
[56] Meta Transparency Center
[57] DSA recital 82

the new regulation on the transparency and targeting of political advertising."[58]

In opening formal proceedings against X, the European Commission has put into question the platform's "effectiveness of measures taken to combat information manipulation on the platform"[59]. Although the investigation press release refers specifically to the use of Community Notes, we believe that there is a wider concern related to X as a result of their recommender system's engagement with and amplification of certain political content over other similarly situated political content, which could amount to information manipulation or encouraging virality of specific content on the platform, thereby posing a risk to the integrity of electoral processes.

Even prior to Twitter's purchase and rebranding to X, the company's engineers were worried about certain political content being amplified on the platform[60]. In 2021, they conducted an internal study[61] examining algorithmic amplification of political content on Twitter and found among others that:

1. Political content is systematically algorithmically amplified as compared to if it is shown in a chronological timeline;
2. Tweets posted by accounts from the political right receive more algorithmic amplification than the political left;
3. Right-leaning news outlets see greater algorithmic amplification on Twitter compared to left-leaning news outlets.

The authors, who worked in collaboration with the University of Cambridge, the University College London, and the University of California, Berkeley concluded that "across seven countries we studied, we found that mainstream right-wing parties benefit at least as much, and often substantially more, from algorithmic personalization as their left-wing counterparts" and note that their methodology might also help studying algorithmic amplification of other types of content such as misinformation, manipulation, hate speech and abusive content.

X reference this study in their RA Report, but in doing so, present its findings as evidence of "the difficulty of determining a definitive causal effect of recommender systems and political bias on social media platforms without considering a wider range of intervening variables"[62], without providing further evidence or additional research to support this conclusion. Though this one study is not indicative of the entire landscape and more broadly, researchers are continuing to explore the impact of asymmetric amplification of content on civic discourse, it would be pertinent for X to provide additional information that contributed to the formulation of their conclusions.

---

[58] Commission guidelines under the DSA for the mitigation of systemic risks online for elections

[59] Commission opens formal proceedings against X under the Digital Services Act

[60] According to Twitter's website, "the ML Ethics, Transparency and Accountability (META) team's mission, as researchers and practitioners embedded within a social media company, is to identify both, and mitigate any inequity that may occur". In that context, they consider "algorithmic amplification [to be] problematic if there is preferential treatment as a function of how the algorithm is constructed versus the interactions people have with it."

[61] Belli (2021) Examining algorithmic amplification of political content on Twitter, Twitter ; Huszár F., Ktena S.I., O'Brien C., Hardt M. (2021) Algorithmic amplification of politics on Twitter, Proc. Natl. Acad. Sci. U.S.A.

[62] X (2023) p.55

Though X provides a useful example to illustrate this problem, it is not the only platform that may create risks for civic discourse and electoral processes by means of asymmetric amplification of civic discourse. For instance, there is "clear systematic bias in how the political ads of different parties are delivered" on Meta platforms in Germany, according to a large-scale study conducted at LMU Munich University[63] that discovered significant discrepancies in the cost effectiveness of advertising and the degree to which micro-targeted online ads reached their intended targets.

Just like other underreported risks, asymmetric amplification of content on civic discourse is not user-related but stems directly from the platforms' own design decisions. As we have seen in most recent elections in Europe (notably Romania, the European Parliamentary elections and now in Germany), this risk bears tremendous potential for political misuse and negative effects on civic discourse and electoral processes. As such, it is crucial that in the future risks to electoral integrity be assessed specifically in relation to the design of platforms, whilst being mindful of the free expression considerations in such a case.

3. Verifiable claims harbour trust in compliance and trust with users

As is mentioned on several instances throughout this brief, the way in which mitigation measures and their claimed effectiveness have been presented in the RA Reports that we analysed poses a fundamental problem: they are not verifiable. The contextual information and data that accompanies descriptions of mitigation measures across all platforms analysed are abstract and for the most part unsubstantiated by either qualitative or quantitative evidence, thereby rendering the claims meaningless.

In addition to this, of the data that is provided in the reports that we reviewed, the vast majority was already public, which leads us to the conclusion that a significant part of the content in the RA Reports is not the result of new assessment processes conducted in light of DSA requirements. This ultimately undermines trust in compliance, as it raises the concern that platforms will not conduct new and appropriate assessments to address evolving risks, but merely reference existing data and policies, which extensive research has shown to fall short of mitigating systemic risks.

Thus far, platforms have given two primary reasons to justify this level of abstraction in their reporting. First, they claim that the data is often sensitive, either because it pertains to trade secrets, or because it may inadvertently help bad actors. Second, they claim that the DSA does not explicitly require platforms to submit additional contextual information and data to support their claims about the choice and effectiveness of their mitigation measures. We challenge both of these claims.

[63] Bär D. et al. (2024) Systematic discrepancies in the delivery of political ads on Facebook and Instagram, PNAS Nexus

First, there are ways in which data can be provided without raising the concerns platforms have indicated. To demonstrate this, below we include a table with examples of the disclosures necessary to assess the effectiveness of mitigation measures, which would not necessarily result in "the disclosure of confidential information [...], cause significant vulnerabilities for the security of [a] service, undermine public security or harm recipients"[64]. Where we believe platforms could have a reasonable claim that the data is sensitive we have colour-coded the text in green and provided further context.

In addressing the claim that providers are not legally required to provide substantiating evidence in the RA Reports, we find this approach problematic and misguided. DSA rec. 40 states that one of the objectives of the regulation is to guarantee the "safety and trust of the recipients of the service". Trust is not possible without proof, especially when independent research and leaked documents cast serious doubts regarding many of the claims made in the reports[65]. We believe that this is a unique opportunity for VLOPs and VLOSEs to demonstrate to both users and regulators that they are striving to achieve the core purpose of the DSA, which is to foster a safer, more transparent and trustworthy online environment in which potential societal concerns that may stem from the systemic risks are adequately addressed. Ultimately, providers claim that their platforms are safe and good for society – now is the time to prove it.

Examples of the disclosures necessary to assess the effectiveness of mitigation measures

Type of mitigation measure (as per art. 35 of the DSA)	**Disclosures necessary to assess effectiveness of the mitigation measures**

Adapting the design, features or functioning of their services, including their online interfaces and user control tools	Interface Design
	
	• Deceptive Design:
	  ○ How does the platform define deceptive or manipulative design (per Article 25 of the DSA)?
	  ○ Who does the internal training, oversight, and auditing of how product designers apply

[64] DSA Article 42(5)

[65] For example: 7amleh (2024) Palestinian Digital Rights, Genocide, and Big Tech Accountability ; EU Disinfo Lab (2024) What is the Doppleganger Operation? ; AI Forensics (2024) Pro-Russian Ads Campaigns Approved by Meta from May 1 to May 27, 2024 in Italy, Germany, France & Poland ; CCDH (2025) YouTube's EU Anorexia Algorithm: How YouTube recommends eating disorder videos to young girls in Europe ; Horwitz J. & Seetharaman D. (2020) Facebook Executives Shut Down Efforts to Make the Site Less Divisive, The Wall Street Journal ; Allyn B. & Kerr D. (2024) TikTok executives know about app's effect on teens, lawsuit documents allege, NPR

	deceptive design standards? ○ What user behavior or user experience signals does the company use to measure design deceptiveness and manipulation? [66] ● Impact of the design on the exposure of users to content related to systemic risks: ○ Does the design reduce the total number of exposures to risky content? If so, what is an estimate of the number of exposures that are eliminated? ○ Does the design reduce the number of exposures related to platform systems, rather than user choices? If so, how does the distribution of exposures change? ○ Does the design reduce the impact of the severity of the harms related to the exposure of the risk? If so, how is the severity of the harm measured at the platform, and how is that impacted by the design/functionality? <u>User Controls</u> ● How does the company define “directly and easily accessible” user control tools? (Article 27(3)) ● Do user control tools offer a combination of granular controls (e.g. over individual pieces of content) and coarser control over the inclusion of specific topics (e.g. political content)? ● How many users have modified their settings away from the default for the relevant user control? ● What is the impact when users edit their settings away from default? What fraction of users are exposed to systemic risks for each setting option? ● What data or what other measurements are used to assess the effectiveness of user control tools (e.g. behavioural data, qualitative data, user survey data)? ● What user behavior or user experience signals does the company use to measure the effectiveness of time management tools (e.g. time spent on platform/time of day, qualitative data, user survey data, etc)? What are the results? ● What is the distribution of daily time spent for all users? Has this number increased or decreased in the reporting period?

[66] This is an area where high level descriptions of signals is not problematic, but exact signal definitions might pose a problem in relation to trade secrets.

	Platforms should disclose: ● statistical data showing how many individuals use control tools, and ● user survey results showing how useful/effective these tools are.
Testing and adapting their algorithmic systems, including their recommender systems	Definition of main parameters and criteria should include (Article 27(1)-(2)): ● *Input data*[67]: All the sources of raw information used in ranking should be disclosed. This could include item content and metadata, engagement history data, user survey data, quality feedback from users, annotations from raters, user settings, profile and social graph data, context data (day, time, location, etc.) and others. ● *Values and their weights (or quartiles):* Platforms should report the complete list of values and their weights to auditors. Because weights are difficult to interpret numerically, and could be claimed by some parties to be trade secrets, the quartile of the weight could be reported publicly instead of raw numeric weights. Auditors should assess weights in relation to mitigation. ● *Metrics*: Platforms should reveal metrics used to measure each systemic risk resulting from recommender systems. For input data and weights, platforms should disclose the information as it applies across the entire user base, as well as with respect to individual user segments reflecting specific age, region, or other cohorts for which platforms specifically tailor their recommendations. The disclosure of weights indicates which signals and predictions have the most influence in the recommenders received by different user segments.

[67] Examples of transparency from companies include Twitter making their source code fully public, including the weights (although it has not been updated in years and is therefore likely out of date) ; Meta also provides a list of engagement signals they use.

	If the platform classifies content for the sake of their policies (e.g. downranking unverified content or promoting authoritative sources), it should reveal: ● what content classifiers are used; ● on what basis (signals, features) is content classified for the sake of their policies; ● how accurate are the content classifiers (precision recall type numbers); ● What is the process for updating and evaluating the criteria? All very large online platforms should reveal: ● What is the prevalence of content related to systemic risks in their recommendation surfaces? ● How many users are exposed to content related to systemic risks on their recommendation surfaces? ● Across the user base, what is the distribution of content related to systemic risks coming from within the user's network versus outside the network? ● How do signals used in recommender systems respond to content related to systemic risks? For example, as a function of the ranking signals value, how does the prevalence of violating content change? Last but not least, very large online platforms should publish comprehensive lists of highly disseminated content, where views, comments, or other engagements have exceeded reasonable thresholds.
Awareness-raising measures and adapting their online interface in order to give recipients of the service more information	How does the company promote such resources? How many individuals have accessed them in the reporting period? Can the company show feedback from the users on the usefulness of their resources? For labels (prompts) added to borderline/sensitive content (such as "click here to learn more about the US election"): ● For content that is found to be violating policies related to the risk, what fraction of exposures on that content is accompanied with the label (prompt)? What is the engagement with the label (prompt) in those situations?

	For resources/content promoted in editorial surfaces (such as "Breaking News Shelf" on YouTube or "News Tab"/ "Election Tab"/"Covid Tab" on Facebook): ● How many users have engaged with the surface in the reporting period? ● What fraction of all exposures to content on the subject covered by the editorial surface occur in this editorial surface? (For example, what fraction of all covid related content on Facebook is viewed in the "Covid Tab"? What fraction of news content on YouTube is seen in the "Breaking News Shelf"?
Targeted measures to protect the rights of the child, including age verification and parental control tools	How is user age assessed by the platform? How is the classification of age-appropriate content done? [68] How accurate are the classifiers when assessing age-appropriateness? How often do people below 18 see age-inappropriate content? How often would they see it without the classifiers? What is the distribution of daily time spent for users known to be under 18? How many users known to be under 18 use time management tools, if provided? How effective are time management tools in reducing unwanted or excessive use? For all relevant settings in the parental controls: ● What fraction of teens, and what total number, are using non-default settings? What fraction of teens had a parent evaluate their settings? ● For each relevant control setting, what is the prevalence of violating content? What is the prevalence of content that the platform deems inappropriate for children? How many teens are exposed to violating content? ● For each relevant control setting, what is the distribution of time spent (per certain time period) for teens? What fraction of time spent is during "sensitive hours", such as typical school hours or sleeping

[68] At a high level, we believe this information is not sensitive, though more granular data on the exact formulas, exhaustive list of signals, etc. might be reasonably redacted in the public version of the RA Reports.

	hours?
For all mitigation measures	Companies should define and reveal: success criteria, test methods, test results, and their conclusions. Broadly speaking, we should expect mitigation measures to impact the scale, cause, or nature of the risks: ● Scale: How many people are exposed to the risk? How many total exposures over some reasonable time period e.g. a month?[69] ● Cause: How many exposures are due to platform recommendations or platform design, rather than user choice? (i.e. violating content recommended to the user vs. violating content from an account that the user proactively chose to follow, or violating content dm'ed to a user from a stranger vs. violating content dm'ed to a user from an account they proactively chose to follow) ● Nature: How concentrated are the exposures? How many exposures occur in vulnerable populations? How harmful is each individual exposure? The platforms should, for each mitigation measure, give some quantitative estimate of the impact of the mitigation measure on the scale, cause, and nature, or at the very least whichever risk dimension is most relevant to the mitigation measure. For example, if the main impact of the mitigation measure is to reduce the scale of the risk, then they need to estimate how many exposures this measure has eliminated. Platforms should specify limitations having a direct bearing on the efficacy of a mitigation measure such as partial geographic rollout across EU Member States.

[69] TikTok and YouTube make this public, specifically for policy violating misinformation, in their Code of Practice on Disinformation Reports.

4. Comprehensive stakeholder engagement

The DSA Civil Society Coordination Group, along with the Recommender Systems Taskforce and People vs Big Tech collectively count nearly 200 global, local and European organisations and academic researchers with a broad range of subject-matter expertise. Of all the organisations represented, none were consulted in the process of conducting the Risk Assessments, nor during the subsequent drafting of the RA Reports.

On two occasions, civil society organisations and VLOPs/VLOSEs had the opportunity to have an exchange during events organised by the Global Network Initiative (GNI) in collaboration with the Digital Trust and Safety Partnership (DTSP)[70]. Though these events were occasions during which civil society representatives were able to understand how providers approached compliance with the DSA in the context of the RA Reports, they were not spaces in which to meaningfully exchange on the substance, nor was any feedback or insights reflected in any RA Reports.

In accordance with DSA recital 90, VLOPs and VLOSEs should engage in a consultation process with experts to conduct their risk assessments and design their risk mitigation measures. It would be pertinent, and logical therefore, for RA Reports to subsequently detail how this process was conducted.

In the reports across all companies that were analysed, there is little to no mention of “the best available information and scientific insights”, while very few external links that cite independent research are featured. It is also not clear how the assumptions were tested with “the groups impacted by the risks and measures” taken by the platforms. Overall, a common theme we saw in the RA Reports is that there was not thorough information about which civil society organizations or experts were consulted and utilised, how those collaborations were implemented into the mitigations or analysis, how they produced fruitful outcomes, and what those fruitful outcomes were.

We hope that this brief showcases that many of the shortcomings in the RA Reports could have been prevented, had there been meaningful stakeholder consultation during the process of conducting risk assessments.

It is impossible for any company to have all the in-house expertise necessary to evaluate all the possible systemic risks that its services may pose, especially while these risks are evolving. It is also impossible to implement effective mitigation measures for all risks without involving affected communities and external experts for that specific purpose. It is therefore imperative that platforms undertake stakeholder engagement as a comprehensive, multifaceted process that involves systematically consulting and listening to subject-matter experts and impacted communities, notably those most at-risk.

[70] European Rights & Risks: DTSP & GNI Stakeholder Engagement Forum

Similarly, third-party research must be consulted systematically in order to best inform risk assessments. The insights drawn from the varied research available on key issues directly related to systemic risks should be reflected in the RA reports and – crucially – the platforms must be clear and specific in indicating which piece of research or consultation process has informed which of their statements[71]. Publishing some types of feedback logs or consultation summaries would also be particularly insightful.

This initial feedback by CSOs seeks to demonstrate how independent expertise can serve as a complementary force in improving the assessment and mitigation of systemic risks with each iteration.

[71] For example TikTok (2023) page 18, though still more specificity would be required.

Conclusion

The initial analysis of the first round of Risk Assessment Reports under the DSA highlights both useful practices and significant gaps in how VLOPs and VLOSEs identify, assess, and mitigate systemic risks. Based on the reports examined, we conclude that they fail to adequately assess and address the actual harms and foreseeable negative effects of platform functioning. Without thorough risk assessments, appropriate mitigation measures cannot be determined, and the unsubstantiated nature of the claims on mitigation measures raise concerns about compliance and undermine trust. To ensure future iterations of these reports advance the public interest, foster trust, and demonstrate effective compliance, the following recommendations are proposed:

1. Address Platform Design Risks

- **Focus on Design-Related Risks:** Future RA Reports must focus more thoroughly on risks stemming from platform design, particularly recommender systems, which amplify harmful content and contribute to systemic risks such as mental health issues and political polarization.

2. Enhance Transparency and Data Disclosure

- **Provide Verifiable Data:** Platforms must disclose quantitative and qualitative data in order to substantiate the effectiveness of mitigation measures. This includes metrics on exposure to harmful content, user engagement with control tools, and the impact of design changes. Users want to trust platforms – but there cannot be trust without proof.
- **Improve Reporting Formats:** Some platforms have effectively adopted digestible formats with clear visualizations, rubrics, and metrics. This improves readability and comparability across reports and should be adopted by all platforms to enable regulators, civil society, and researchers to better evaluate platform compliance and the effectiveness of mitigation measures.

3. Engage Meaningfully with Stakeholders

- **Involve Civil Society and Experts:** There is a growing wealth of knowledge generated by experts around the world. To make use of this, platforms should systematically consult with civil society, researchers, and impacted communities during risk assessments and mitigation design, as required by DSA Recital 90. Meaningful stakeholder engagement is essential to ensure that risk assessments reflect the actual harms experienced by users and in building trust with affected communities. Simply repackaging existing stakeholder engagement that happens at global scale is not enough to address DSA-specific risks.
- **Incorporate Independent Research:** RA Reports should reflect insights from external studies and provide clear links between research findings and platform actions.

Independent research is critical to identifying gaps in risk assessments and ensuring that mitigation measures are evidence-based.

By implementing these recommendations, VLOPs and VLOSEs can better comply and align with the DSA's objectives, enhance user trust, and contribute to a safer, more transparent digital environment. We hope that this brief clarifies our expectations for how meaningful transparency can be achieved through this iterative exercise, but also to assist the European Commission as they assess platform compliance with their risk assessment and mitigation obligations – it can also provide reflections to complement existing enforcement actions, such as the numerous Requests for Information and investigations. Civil society and researchers remain committed to supporting this process through ongoing analysis, feedback, and collaboration.

This brief was developed by the DSA CSO Coordination Group, an informal coalition of civil society organisations, academics and public interest technologists that advocates for the protection of international human rights, respect for the rule of law and human rights due diligence in the development, implementation and enforcement of the EU Digital Services Act.