Oversight Board

WHY FREEDOM OF EXPRESSION MUST BE THE CENTERPIECE OF SYSTEMIC RISK ASSESSMENTS

THE OVERSIGHT BOARD, JUNE 2025

Executive Summary

In 2022, the European Union approved the Digital Services Act (DSA), legislation promising to protect user rights and placing a regulatory requirement on platforms to identify and mitigate risks resulting from their online services. Crucially, the DSA stipulates that online platforms, including social media companies, must "give particular consideration" to freedom of expression when deciding how to address the serious harms to society identified under this framework. Since these platforms published their first assessments in late 2024, several challenges to this aim are becoming apparent, some derived from the ambiguity of the DSA's key terms, others from missed opportunities to integrate global human rights standards into these assessments.

Building on the work of many organizations active in this field, the Oversight Board believes it is crucial that human rights, particularly freedom of expression, are placed at the core of systemic risk assessments. In that spirit, this paper sets out four focus areas that could help to enhance platform accountability and improve how content is governed, as part of a consistent and effective rights-based approach:

- **Clarify the meaning of systemic risks.**
  Ambiguity over this DSA term could leave the door open for overbroad interpretations, potentially incentivizing restrictions on speech.
- **Draw on global human rights standards.**
  Fully integrate such standards across all categories of risk assessment for more consistent reporting. Mainstreaming global human rights is more effective than treating them as a standalone category.
- **Embed stakeholder engagement into identification of risks and design of mitigations.**
  By following the practices set out in the UN Guiding Principles on Business and Human Rights (UNGPs), platforms can more meaningfully show how stakeholder engagement shapes their responses to risk.
- **Deepen analysis with data.**
  Quantitative and qualitative data are equally valuable to reporting. Companies should more openly use appeals data supported by insights from external oversight mechanisms to show whether mitigations are effective in respecting freedom of expression and other human rights.

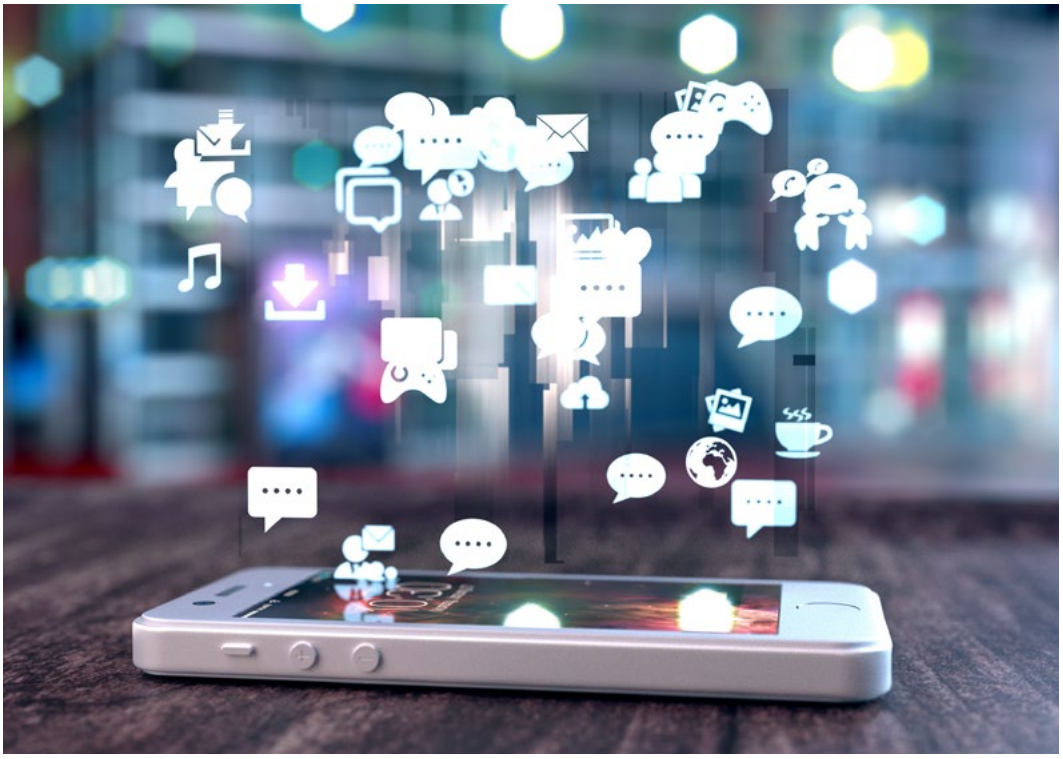
Introduction

Recent EU regulation of online platforms introduces a new, risk-based approach to online services, focusing on how platforms may create or amplify certain types of harm. The DSA seeks to regulate social media to establish "harmonised rules" for a "trusted online environment" in which human rights are respected. It requires "very large online platforms" (VLOPs) to disclose the steps they are taking to prevent their services from harming people and society. The early "systemic risk assessments" published by VLOPs provide insights into how platforms identify, evaluate and mitigate risks, including to human rights, arising from the design and use of their systems, as required by DSA Articles 34 and 35. Although the DSA has the potential to enhance transparency and support human rights, the incentives it creates could also lead to excessive restrictions on freedom of expression globally.

Reconciling Risk Mitigation and Respect for Freedom of Expression

Many of the risks the DSA addresses reflect the issues the Board has prioritized in its cases. For example, the DSA (Recital 86) requires platforms to "give particular consideration to the impact on freedom of expression" when choosing how to mitigate systemic risks. This consideration is closely linked to the Board's mandate, which centers on ensuring respect for freedom of expression and identifying when speech restrictions may be justified to protect other rights or interests. Our decisions, which are binding on Meta, tackle the most challenging content moderation issues, and examine how Meta's policies, design choices and use of automation impact people's rights. These decisions provide insights into how to reconcile the identification and mitigation of risks on Meta's platforms with respect for freedom of expression and other human rights.

The Board emphasizes that systemic risk assessments must include greater focus on respect for human rights, including freedom of expression, if they are to enhance meaningful platform accountability to users and improve content governance in line with the DSA's objectives. This is consistent with recent work produced by organizations – including Global Network Initiative (GNI), Digital Trust & Safety Partnership (DTSP), Access Now and the Center for Studies on Freedom of Expression and Access to Information (CELE) – and other experts across the field, to deepen understanding of systemic risks, anchoring risk assessments in global human rights standards, and highlighting potential threats to freedom of expression and risks of political interference. Drawing on this work and its close analysis of the first systemic risk assessments, the Board offers the following reflections.



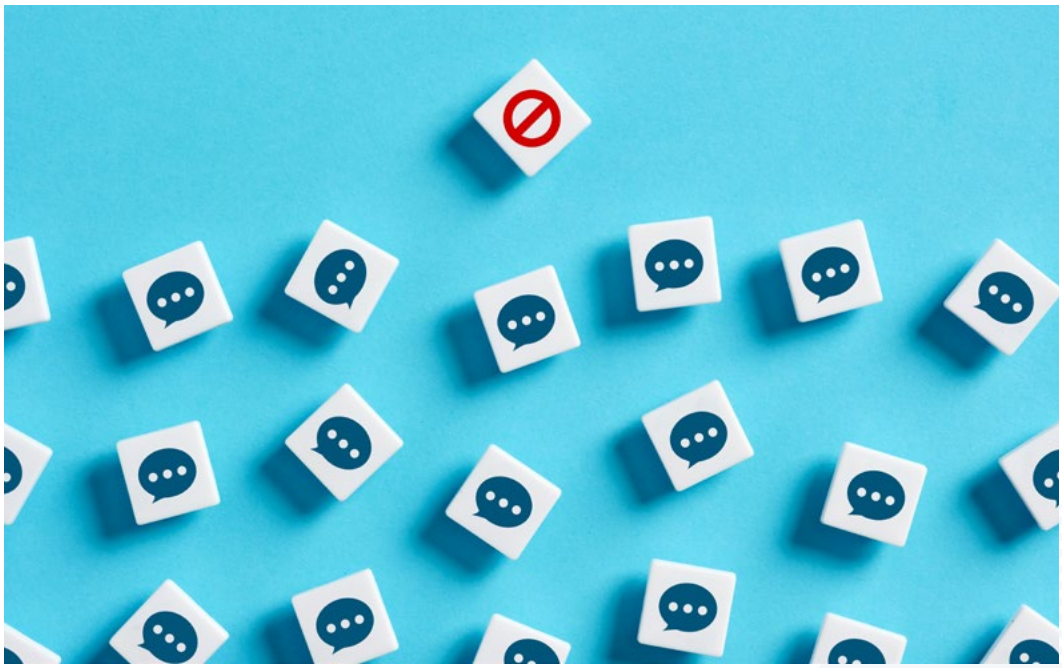
Clarify the Meaning of Systemic Risks

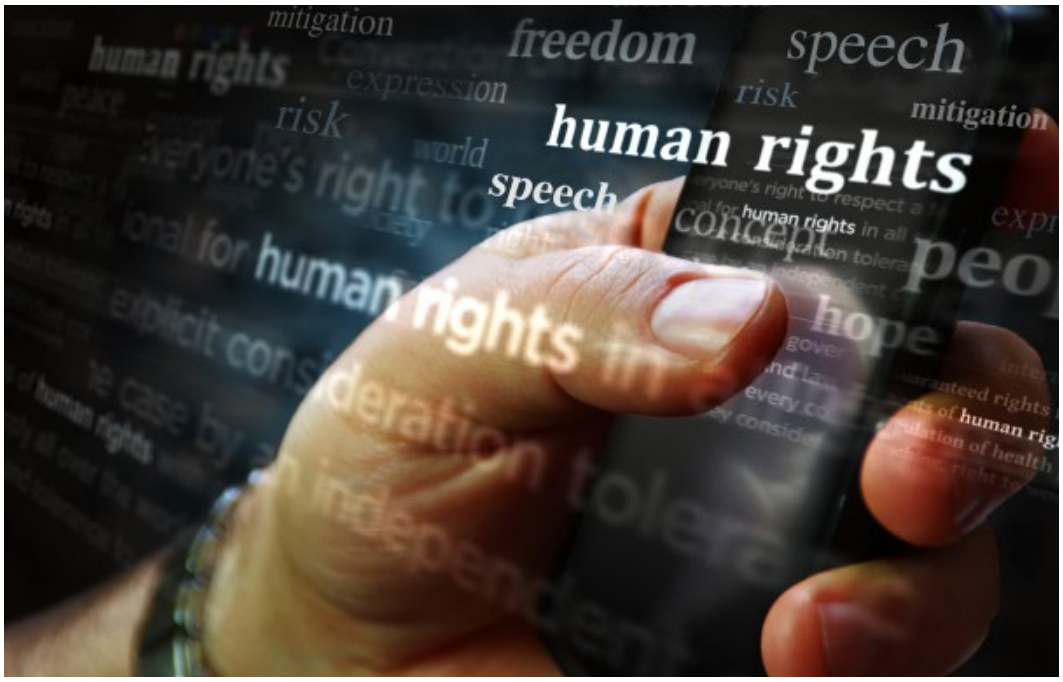
The first reports are limited by the lack of a shared understanding of what the term "systemic risks" means. It is not defined in the DSA and is not rooted in global human rights law. While the Board acknowledges the DSA's deliberately flexible approach of allowing the meaning to develop over time, this shifts the responsibility over to platforms to thoughtfully interpret the concept. Given this, it is understandable that platforms often default to a narrow, compliance-focused approach, which can hinder a meaningful understanding of systemic risks developing. The result is the reduction of systemic risks analysis to a checklist exercise, as largely seen in the initial publication of platforms' risk assessments in 2024.

Most platform reports refer only to the DSA's listed systemic risk categories ("illegal content," "negative effects" on "fundamental rights," democratic processes, public security, "gender-based violence" and the protection of minors) and its 11 mitigation measures (e.g., "adapting" and "adjusting" design choices and "recommender systems"). Platforms are largely silent on whether their assessments identified new risks or led to the rollout of new mitigations, and do not challenge presumed connections between their platforms and specific risks. This ambiguity, in turn, may facilitate platforms missing or obfuscating new threats and emerging trends.

Incentivizing Speech Restrictions

From a freedom of expression perspective, ambiguity over the term's meaning may lead to overbroad interpretations and arbitrary enforcement, incentivizing excessive restrictions on speech. This could stifle diverse opinions and potentially chill platforms' commitments to providing spaces for open discourse on challenging and sensitive topics. Consequently, this could deter users' freedom to express themselves on these platforms. It has also the potential to undermine some of the benefits that the DSA may bring in terms of greater access to user remedy and increased transparency.





Draw on Global Human Rights Standards for Systemic Risk Reporting

The DSA treats human rights as a standalone category rather than integrating it across risk areas, leading to fragmented approaches on how platforms identify, assess and mitigate risks. This is especially problematic given the DSA's novel standard that mitigations must be "reasonable, proportionate and effective," which lacks clear implementation guidance. By placing human rights in a standalone category, the DSA misses the opportunity to integrate human rights considerations comprehensively into systemic risk governance. This prompts platforms to prioritize certain rights over others and discourages them from assessing how each risk area or "influencing factor" may affect human rights as a whole. Recent research from the CELE, an Argentina-based NGO, argues that the risk-based approach "pushes rights out [from] the center stage of Internet governance and may create a logic of 'symbolic compliance' where [the] governance role of rights is further diminished." Drawing on global human rights standards could support a more consistent and rights-based approach to systemic risk reporting, helping align methodologies while ensuring a common framework for assessing impacts on rights.

Nuances Overlooked

This fragmented treatment becomes particularly evident in the context of freedom of expression. While standalone reporting may cover concerns about content moderation practices, account suspensions or misinformation, it often overlooks more nuanced issues. For example, it may fail to consider how other risk areas like “illegal content” or "influencing factors” like automated detection, recommendation algorithms or search functionalities can have systemic impacts on freedom of expression, even when these effects initially seem limited. Or, in another instance, when platforms cooperate with governments on content takedowns, it is often unclear how such requests are made, recorded or acted upon.

This lack of transparency has been a recurring issue identified in the Board’s case work, which has examined the opaque and inconsistent nature of state requests (see Shared Al Jazeera Post, UK Drill Music and Öcalan’s Isolation decisions), and their potential to suppress freedom of expression. Platforms also rely heavily on automated systems to detect and remove content, which can, on the one hand, lead to the overenforcement of political and counter speech. On the other, reducing reliance on automation can also carry risks, with uneven consequences for different users. The Board recently recommended that Meta examine the global implications of its decision, announced on January 7, 2025, to reduce reliance on automation for some policy areas.

Mainstream Human Rights

To mainstream human rights as a cross-cutting issue, platforms could benefit from greater clarity and implementation guidance on how to identify and assess risks through a rights-based framework with clear and consistent criteria. While many platforms have developed their own approaches, they often reference a variety of frameworks in their reports, from the UNGPs to risk models from unrelated fields like finance and climate change. This leads to inconsistent evaluation of factors such as scope, scale, irremediability and likelihood of potential adverse impacts. All this hinders the ability of stakeholders to compare risks across services, and assess industry-wide harms and limitations on users’ abilities to speak freely.

Drawing upon guidance from international treaties and the UNGPs could help ensure that efforts to identify and assess systemic risks do not unduly infringe on human rights. The UNGPs offer a structured approach for assessing human rights impacts, emphasizing stakeholder engagement, context and attention to vulnerable groups. They involve well-established guidance on evaluating the scope, scale, irremediability and likelihood of potential adverse impacts on human rights.

Using the UNGPs would enhance cross-platform comparability and ensure that risk assessments go beyond what is immediately visible or quantifiable, capturing broader and longer-term impacts embedded in platform design and operation.

Distinguish Between Risks and Mitigation Measures

To navigate these challenges, platforms also need a structured way to distinguish between prioritizing risks and determining mitigation measures. A rights-based approach could help platforms apply carefully calibrated measures, rather than oversimplifying assessments based on risk prioritization. This approach should include an evaluation of the impacts of mitigation strategies themselves, using clear, rights-specific criteria. For example, measuring the effectiveness of content moderation would require assessing content prevalence, volume of decisions, enforcement error rates and appeal outcomes. This would ensure that responses to risks do not generate new or disproportionate impacts, while resulting in more granular transparency and access to data to support third-party research into moderation trends.

While the DSA aims to establish a framework for evaluating mitigation measures by requiring them to be "reasonable, proportionate and effective," it lacks clear implementation guidelines. As with risk identification and assessment, this leaves much to the discretion of platforms and results in the use of divergent methodologies, which can affect the quality, effectiveness and timeliness of these mitigations.

Clearer guidance on how to evaluate and implement mitigation measures could be achieved by drawing on existing global frameworks for evaluating restrictions on speech: namely, the three-part test for legitimate restrictions on freedom of expression, based on Article 19 (3) of the International Covenant on Civil and Political Rights (ICCPR), and its relevance to companies under the UNGPs. This would allow platforms to better evaluate mitigation strategies by integrating speech concerns and other legitimate aims. Another benefit would be ensuring that freedom of expression and civic discourse are not treated as a standalone "risk" area, but mainstreamed as a cross-cutting issue.

Organizations That Bridge the Gap

Embracing existing frameworks would challenge assumptions that freedom of expression is always in tension with respect for other human rights and societal interests, and encourage innovative approaches to risk mitigation. This route would also clarify the relationship between the DSA's standard of "reasonable, proportionate and effective" and well-established human rights frameworks, like the ICCPR's Article 19 three-part test of legality, legitimacy, and necessity and proportionality. The Board applies this three-part test in all our cases to assess whether Meta's speech interventions meet the requirements for legality, legitimate aim, and necessity and proportionality. This provides a transparent and replicable model for rights-based analysis that platforms can adopt in their own mitigation efforts.

A Consistent, Global Response

Systemic risk frameworks designed under regional regulatory regimes, such as the DSA, could end up shaping regulatory approaches in other regions. Therefore, it is crucial for the regulator to clarify the cross-cutting role of human rights across all risk areas and for platforms to adopt frameworks rooted in global human rights standards to ensure their systems effectively mitigate risks in regional jurisdictions, while maintaining global consistency. As the Board's extensive work demonstrates, relying on global standards requires consideration of local and regional contexts, both when identifying risks and designing mitigations. While harms to individual rights may manifest differently in different regions, applying a global framework can ensure that a company's response is consistent and grounded in respect for freedom of expression.

Embed Stakeholder Engagement into Assessments and Mitigation Design

Although all platforms refer to stakeholder engagement (such as civil society, academia and marginalized communities) in their reports, there is limited insight into how this input informs systemic risk assessments. While platforms set out their consultation processes in detail, they do not clearly draw connections between the outputs of those consultations and their analysis of risk or evaluation of mitigations. This reporting on stakeholder engagement also fails to align with the good industry practices outlined in the UNGPs. Specifically, with the lack of clarity on how engagements are structured, which stakeholders are involved and what concerns are raised, it is difficult to understand how stakeholder insights influence platforms' responses to individual risks, before and after mitigations are applied.

Diverse Perspectives

Meaningful stakeholder engagement should prioritize the input of individuals and groups most affected by platform decisions by actively seeking expertise and diverse perspectives. Moreover, this type of engagement is essential for considering regional and global factors when assessing systemic risks and mitigations. While the DSA emphasizes localized risk assessment, current methodologies often fail to account for local diversity (e.g., the EU's different languages and cultures), since platforms mainly focus on structural issues affecting their systems. This is exacerbated by a lack of targeted stakeholder engagement, leading to risk assessments that fail to capture the complexity of local contexts.

The Board's prioritization of stakeholder engagement in cases and policy advisory opinions highlights how such efforts can increase transparency and participation, and amplify the voices of people and communities most impacted by platform decisions (see the "Shaheed" policy advisory opinion). Additionally, the work of expert organizations, such as the Global Network Initiative and Digital Trust & Safety Partnership forum, underline how multi-stakeholder consultations with diverse experts can enrich both risk assessments and mitigation strategies, and help platforms align these processes with a rights-based approach.

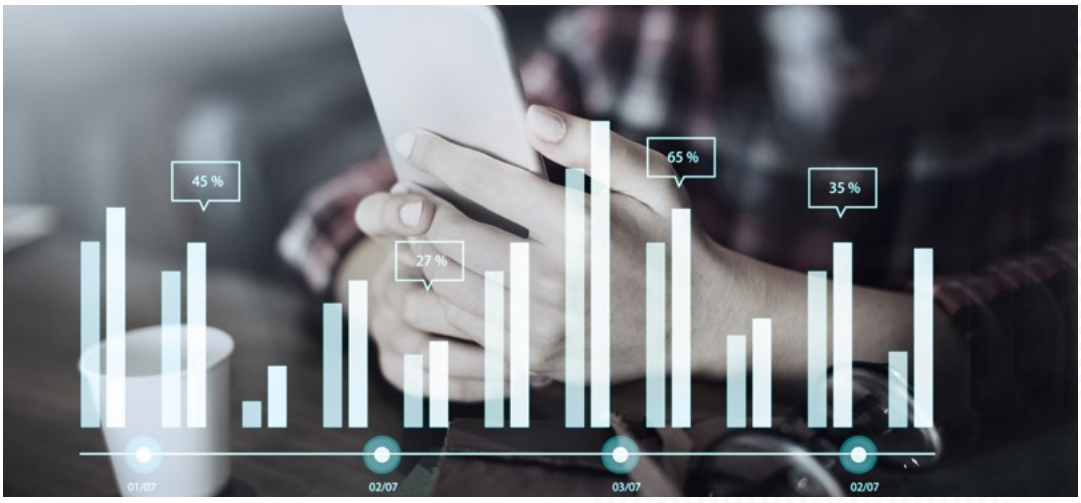
Deepen Analysis with Appeals Data

Since the first reports by platforms are primarily qualitative, they provide limited insight into the quantitative data used to assess risks and mitigation measures. When cited, metrics are often high level and duplicate pre-existing transparency report disclosures. Building on the Board's experience, one way to evaluate the effectiveness of mitigation measures, particularly on freedom of expression and other human rights, is to draw on both qualitative and quantitative assessments of user appeals data, such as on decisions to remove or restore content. Appeals are not only a mechanism for error correction, they are also a vital safeguard for protecting free speech by revealing which enforcement practices may be suppressing lawful expression. User reports and appeals against decisions to leave content online can also highlight where enforcement practices may be failing to properly curb harmful content.

Enforcement Trends as Indicators of Risks

Appeals can also offer valuable insights into enforcement accuracy and residual risks. For example, data on appeals volume, geographic location, relevant policies, associated risk areas and outcomes can help determine which mitigation measures are effective over time – and which require improvement. Receiving hundreds of thousands of appeals annually from around the world, the Board's data could help highlight enforcement trends as potential indicators of risks, such as censorship of journalistic content, and over- or underenforcement of policies during a crisis, as well as help to evaluate the effectiveness of mitigations. This, in turn, could supplement platforms' own processes, contributing to independent oversight.

By systematically analyzing, openly reporting and meaningfully integrating data into risk assessments, platforms will not only enhance the effectiveness of mitigation but also strengthen trust in their commitment to safeguard human rights.



Conclusion

Now the initial rounds of assessments have been published and as platforms develop the next round of reports, the time is right to refine methodologies to ensure that products, platform features and content moderation systems are evaluated with greater precision, depth and robustness. A transparent and multi-stakeholder approach, bringing together diverse expertise and perspectives, is essential to support this endeavor. It is crucial that human rights, particularly freedom of expression, are placed at the center of systemic risk assessments to safeguard speech, rather than to serve as a mechanism for its restriction.

By drawing on its expertise, the Board is committed to help develop rights-based approaches that centrally position freedom of expression. Given the iterative nature of assessments, the Board encourages platforms to incorporate feedback and for regulators to take these insights into account when designing guidance for platforms and auditors.

The Board looks forward to working with interested organizations and experts on systemic risk assessments and mitigation.

Oversight Board