GLOBAL NETWORK INITIATIVE

Digital Trust & Safety Partnership

2026 Rights & Risks Forum

EVENT REPORT

Dublin, April 2026

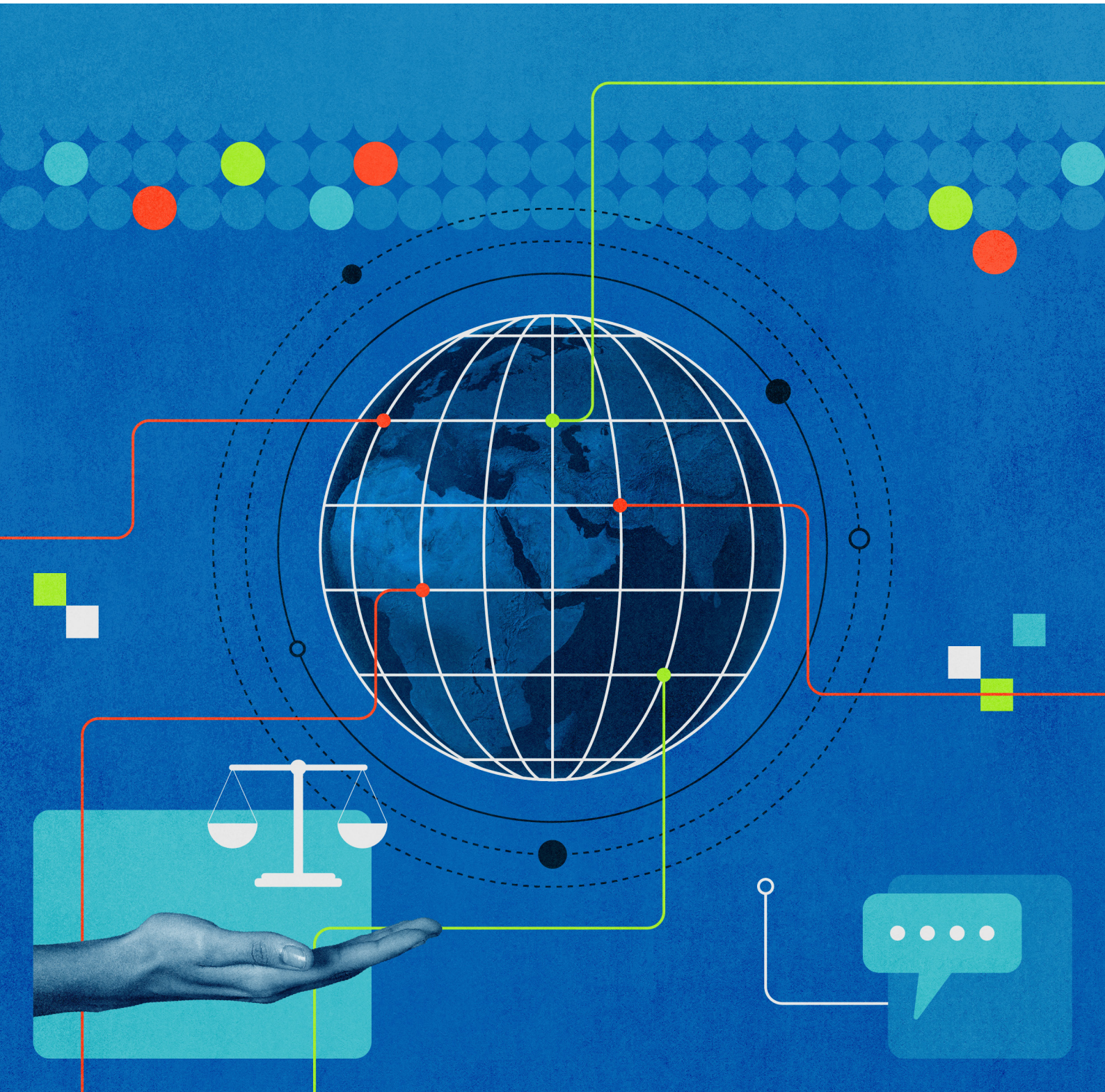


Table of Contents

Executive Summary

The annual Rights & Risks Forum held its fourth edition in 2026. Hosted by the Digital Trust & Safety Partnership (DTSP) and the Global Network Initiative (GNI) in Dublin, it convened industry and civil society experts to discuss and inform the risk assessment practices of online services and platform companies.

The Forum sought to ensure that these companies maintain a strong focus on the protection of fundamental rights, especially freedom of expression and privacy, at the centre of digital safety risk assessment methodology and practice. It focused on how digital services and technology companies can appropriately assess and mitigate risks to fundamental rights, taking into account responsible business conduct guidance, diverse regulatory requirements and expectations, trust barriers, and practical considerations.

Discussions at this year's Forum highlighted how, in the regulatory discourse, contested definitions of systemic risk threaten to eclipse human rights concepts. Against a backdrop of democratic backsliding, participants highlighted that the perception of digital spaces as inherently risky may provide cover for political abuse and government overreach. Participants argued that insufficient attention is currently given to practical implementation of the risk-based regulatory approach. Especially given the proliferation of online safety laws with distinct risk management requirements, shifts in regulatory focus from rights to risks can lead to restrictions or removals of lawful content. Participants also unpacked the ways in which shrinking civic space impacts risk identification and mitigation, and how political developments impact trust between civil society, companies, and governments.

The Forum highlighted a range of areas of harm and harm mitigation, including: image abuse, child safety and child rights, age and identity assurance, monetisation, and AI. In each case, participants discussed the need to understand and approach these issues through safety and human rights frameworks and methodologies, including robust rights and safety assessments, which are integrated into the design and deployment of digital services and embedded in internal governance.

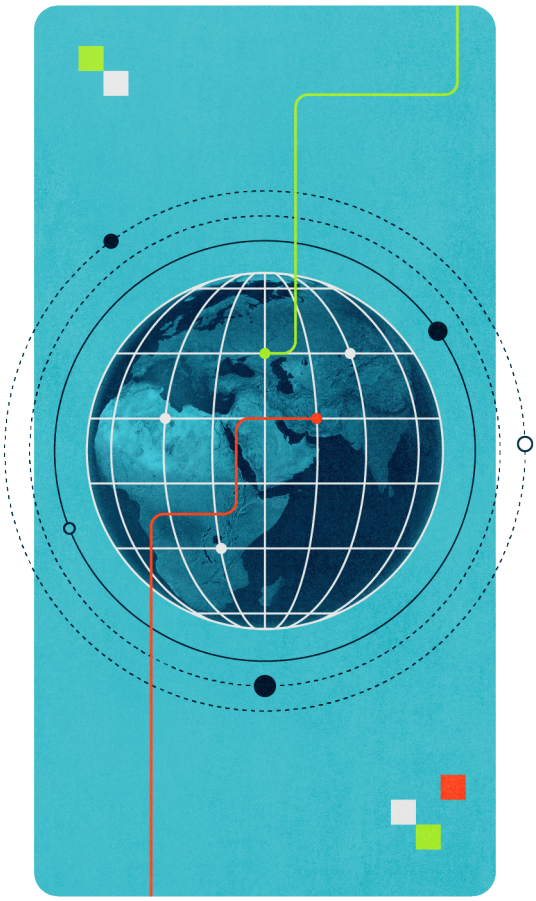
Assessing and mitigating risk across large company operations in a rapidly evolving risk environment is complex and challenging work. This year's Forum highlighted the need for regulatory frameworks to improve the digital environment for users and protect and promote human rights. For company risk assessments and mitigations to be meaningful, they must be well-informed, tested, and take account of specific language and cultural contexts while adhering to universal human rights principles and taking account of the particular challenges in areas like counter-terrorism and child protection. Participants broadly agreed that regulatory risk management requirements should take account of and build on practices grounded in frameworks like DTSP's Best Practices and the Safe Framework (ISO/IEC 25389), the GNI Principles on Freedom of Expression and Privacy, as well as broader, sector-agnostic frameworks like the OECD Guidelines on Multinational Enterprises and the UN Guiding Principles on Business and Human Rights.

This report highlights a number of recommendations drawn from the discussions. These are not recommendations from the organisers, but rather ideas that emerged from the Forum that are relevant for companies, governments and civil society to consider as they work to deliver effective regulation, risk management, and better experiences for people across the world. DTSP and GNI look forward to continuing these conversations.

About the Forum

Background

On 16 and 17 April, the Digital Trust & Safety Partnership (DTSP) and the Global Network Initiative (GNI) hosted the 2026 Rights & Risks Forum in Dublin, Ireland. The Forum's goal is to ensure that the protection of human rights, especially freedom of expression and privacy, is at the center of company approaches to risk management. This was the fourth annual edition of the Forum, the third to be held in person, and the first to be held in Dublin. The focus of this edition was on how tech companies can appropriately assess and mitigate risks to human rights, taking into account responsible business conduct (RBC) guidance, diverse regulatory requirements and expectations, trust barriers, and other practical considerations.



This year's program contextualized rights protection and risk management amidst a changing global environment that is shaping the design and implementation of regulatory frameworks. More than 90 experts attended the Forum. In addition to the organizers, participants were company practitioners from DTSP and GNI member companies, civil society experts, academics and independent researchers, and international organisation experts. Participants came from around the world, with expertise in jurisdictions such as Brazil, Chile, the European Union (including Belgium, Ireland, Germany, Greece, Spain, and the Netherlands), Lebanon and the broader West Asia and North Africa region, Indonesia, Iran, Mexico, Pakistan, Uganda, the United Kingdom, and the United States. In order to encourage candid discussion, government officials and regulators were not invited to the Forum. The organisers deeply appreciate the participants' time and insights.

The Forum was held under the antitrust policies of the GNI and DTSP, the GNI's Code of Conduct and the Chatham House Rule: "Participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed."

The event was hosted and sponsored by Google, and supported by Meta, Bing, and TikTok. Sponsorship helped to cover the costs of hosting the conference, as well as event organization, and civil society travel.

This high-level summary was prepared after the event by GNI and DTSP. It does not necessarily represent the views of DTSP's and GNI's members, nor of the individuals or organisations that participated in the Forum. It seeks to capture key themes, learnings, and recommendations. Prior to publication, we made a draft available to a selected group of participants for a review for accuracy.

Context

The policy conversation on online safety and content regulation has shifted from technocratic problem-solving to high-stakes geopolitical debate, tied up with national trade and security interests and set against a global backdrop of democratic backsliding. At the same time, support for restrictive bans, stemming from concerns for children's safety and well-being, could enable political abuses and undermine democratic discourse and accountability. As online safety laws and mandatory risk assessments proliferate across diverse jurisdictions including Australia, Brazil, the EU, India, and the UK, it is important to consider how effective risk assessments can be properly grounded in established global human rights frameworks.

Goals

Given this context, the organisers articulated the following objectives in advance of the Forum:

- Companies share information about how they assess and mitigate risks to fundamental rights.
- Civil society experts share analysis, questions, and recommendations that could inform rights-based risk assessment and mitigation measures.
- Collectively brainstorm what actors across the field could do to help enable risk assessments to center rights-based approaches and be ongoing learning exercises.

While the Forum represents a key opportunity to shape and deepen engagement between participating companies and their civil society and academic stakeholders, companies are expected to undertake a variety of engagements across different stakeholder groups and types of expertise.

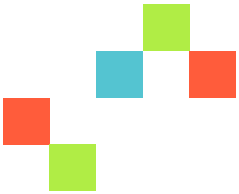
Key Themes, Learnings, and Recommendations

Forum participants discussed how to navigate diverse regulatory expectations and requirements by placing the protection of human rights at the center of company risk management practices. They focused on how to tackle obstacles to civil society engagement in risk management, including the definition, assessment, and mitigation of risks. The conversations were framed by the rapid evolution of AI, which is amplifying the scale and speed of digital harms but also deployed to tackle trust and safety challenges, giving rise to new accountability, privacy, and censorship risks.

1 The Evolving Context of Privacy and Freedom of Expression

1.1 Overall political, social, and regulatory context

In a regulatory landscape characterised by democratic backsliding and the weakening of multilateral and multistakeholder spaces, it is important that democratic values are upheld and that compliance does not become a pretext for censorship. Competing legal and political approaches to freedom of expression complicate international alignment and create negative unintended consequences, including exploitation by authoritarian regimes.



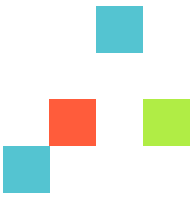
Key Learnings	Recommendations
Democratic backsliding has increased the risks of arbitrary, unnecessary, and/or disproportionate restrictions on users of digital platforms and services.	Companies should actively resist government overreach, and companies and civil society should advocate for independent judicial review of content restrictions and other procedural safeguards. Governments should ensure appropriate transparency, oversight, and accountability of their own activities.
Growing gaps in the public understanding of freedom of expression online are making way for reactive and likely ineffective policy responses.	Governments, digital service companies, and civil society organisations should seek to increase public understanding of digital rights and international human rights standards.
The focus on prescriptive risk assessment threatens to eclipse human rights concepts as the guiding light for technology and platform regulation.	Governments should explicitly center democratic values and international human rights standards within the text, design, and enforcement of risk-based public policies and regulations.
Diverging international approaches to freedom of expression create regulatory complexity and friction.	Companies should base their approaches to risk management in international human rights law and the UNGPs, independently of whether they are complying with legal obligations or conducting a voluntary assessment. They should adopt a 'think global, act local' approach to risk assessment, including consulting with local civil society organisations. Civil society and governments should advocate for universal human rights principles and work to build shared understanding across jurisdictions based on them.

1.2 The transatlantic and transnational debates on safety and content regulation; defining risks and navigating diverse regulatory expectations

Global digital safety debates are increasingly contested and shaped by domestic political and global geopolitical factors. The European concept of systemic risk has come to dominate compliance models globally, despite the fact that it leaves many key questions open to interpretation, and presumes robust regulatory capacity and a judicial system grounded in human rights frameworks. Digital spaces are increasingly framed as inherently risky, giving credence to popular proposals for bans and other restrictive regulations. Civil society stakeholders highlighted examples, including in Indonesia and the Western Balkans, where regulatory frameworks have been used to mandate arbitrary restrictions on users of digital platforms. Risks should be carefully defined, culturally informed, and understood using international human rights principles.

Key Learnings	Recommendations
The risk-based approach has become central to the regulation of digital platforms and services, yet the conceptualisation of risk, and systemic risk in particular, is increasingly contested.	To prevent arbitrary (re)definitions of risk, governments, researchers, and civil society organisations should anchor digital safety frameworks in international human rights standards. Experts in business and human rights should do more to clarify how risk management frameworks can be legally and intellectually grounded in human rights. For example, risk management related to harms to children could be an opportunity to use and refer to the Convention on the Rights of the Child; risk management related to harms of women could be an opportunity to use and refer to the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW).

Authoritarian and backsliding regimes are exploiting regulatory ambiguity to demand censorship.	Governments should be precise in their regulatory requirements, so the text of the law itself clearly protects human rights and does not overly rely on good intentions of the interpreters to be rights-respecting. Civil society should advocate for precise rights-respecting requirements and companies should use their leverage to encourage this precision, as well. Governments should privilege content-neutral mitigation of safety risks.
The export of digital services regulations from stable democracies with independent judiciary, grounded in human rights, to conflict zones, or regions experiencing democratic backsliding, can facilitate the suppression of legitimate political discourse.	Governments should ensure that global perspectives are included in regulatory impact assessments, as tactics for governmental misuse are often tried in one jurisdiction before moving to another. This can help prevent unintended, repressive harms, especially in the Global South.
Imprecise risk definition may lead to the suppression of lawful content.	All actors should seek precise and well-informed risk definitions, rooted in International Human Rights Law, to avoid inadvertent restrictions on legitimate social and political discourse.



2 Stakeholder Engagement In Complex Environments

2.1 Challenges of stakeholder engagement in complex environments; practical responses for higher-risk, low-trust environment

Civil society organisations often find that their real influence on company products and policies is limited. At the same time, superficial or one-size-fits-all stakeholder consultations can inadvertently legitimise authoritarian actors and regimes. Company practices should seek to strengthen, rather than degrade, trust-based networks and safeguard participants from possible retaliation from powerful actors.

It is important to understand that state-imposed restrictions on civil society also limit the ability of companies to operate and uphold human rights. This happens both directly and indirectly: laws restricting association or expression can also be used against companies, and silencing civil society clouds the ability of companies to accurately understand and assess risks, while also enabling and emboldening government overreach.

Key Learnings	Recommendations
Civil society experts have noticed an overall decline in company engagement with civil society over the past few years.	Companies that have pulled back should seek to reengage, and those that have not should continue to engage with civil society in a meaningful and consistent manner. Companies, civil society, and membership organisations should seek to continue to share information about the value of engagement, especially through public resources like case studies. All actors should explore what the possible structural changes could be to encourage companies to take civil society input more seriously.

Formal engagement structures often hit a ceiling where advice is heard but not incorporated into practice.	Companies should ensure that internal human rights experts are empowered, and formalise and diversify engagement with journalists, media, citizens, and international and local advocacy groups. They should seek input at optimal moments so that insights can genuinely be integrated as part of their policy, product and process design, and share information about how input has been addressed.
Poorly conceived stakeholder engagement can become performative exercises that expose civil society to political risks, rather than protecting users from harm.	Companies must recognise the political environment they are operating in, adopting engagement strategies that are tailored to take account of this context. This could include protective measures for civil society experts, like anonymity, in circumstances where they are at risk. Civil society has developed and should continue to highlight and update guidance to help companies engage more effectively. Companies should follow this guidance, to meaningfully involve stakeholders at all stages of their risk assessment processes, and develop engagement frameworks grounded in the UNGPs.
In crisis situations, and situations of major policy or product changes, companies listening to concerns is not enough: direct answers are needed.	Companies should develop relationships with civil society in advance of crises. Once there is conflict or a major crisis, the same civil society experts who are supposed to engage with the companies could be under direct threat. It will be challenging for them to engage; it's easier if there's an existing relationship. Companies should be expected to explain their actions, and advocates should push them to publish detailed case studies on how they navigate complex environments. Governments should encourage good faith efforts and candor about lessons learned.

Human rights arguments frequently fail to move authoritarian governments on digital rights issues.	Companies and advocates should use a range of values-based and economic arguments with authoritarian and backsliding regimes, demonstrating how overbroad regulations and internet disruptions cause direct social and financial harm. Civil society can continue to monitor and produce data and other evidence of repression when it happens.
Civil society organisations can sometimes face backlash from their constituencies if they choose to directly engage with governments or companies that are perceived by their constituencies as not working in good faith.	Companies should be aware of this dynamic across departments, and consistently seek to mitigate acting transparently and in good faith with civil society. Civil society organisations should make strategic engagement decisions based on their constituencies and goals. They may choose not to engage directly; to engage only on some issues, transparently to their constituencies; or to engage via another trusted intermediary.

2.2 The security challenges and complexities faced by human rights and civil society experts

Civic space is shrinking around the world. Civil society experts and human rights defenders commonly face logistical and resource constraints ranging from physical threats to unnotified account suspensions, especially in jurisdictions marked by low rule of law and democratic indicators and/or conflict.

Key Learnings	Recommendations
In volatile environments, civil society and academia are uniquely positioned to identify problems and solutions that internal corporate teams or external consultants might overlook.	Companies should develop policies to identify and protect at-risk human rights defenders, including by maintaining secure, direct, and rapid-response communication channels. Companies should help civil society move beyond reactive advocacy toward proactive co-design, ensuring the lived experiences of marginalized users in conflict zones directly influence technology architecture. Community research combined with movement lawyering to push for, rather than just "included" in, safety-by-design.
Local advocates can face sudden and unexplained account suspensions and censorship, often having no structured way to reach platforms.	Companies should make appropriate, well-resourced appeals/grievance mechanisms available and provide clear guidelines on how to use them. When these channels fail, advocates should leverage international networks, NGOs, multistakeholder bodies like the GNI, and other appropriate mechanisms to communicate with companies and advocate for content and account restoration.

Content moderation in conflict zones can have life-or-death consequences, cutting civilians off from critical information while leaving authoritarian actors unchecked.	Companies should exercise heightened due diligence in conflict affected/high risk contexts and apply rules consistently across all actors. During crises, they should consider newsworthiness allowances and conflict-sensitive moderation policies that protect access to vital information and support the documentation of abuses.
Travel restrictions and funding constraints make it difficult for civil society organizations, particularly in the Global South, to participate meaningfully in dialogue and deliberation with companies and other stakeholders.	To enable civil society participation, companies must design outreach, including relevant events, to be inclusive and properly resourced. When possible, companies should provide funding for participants and assist with practical obstacles such as visa and travel restrictions. Such financial support should be transparent.

2.3 Government and political pressure on company policies and practice; impact on the conduct of mandatory risk assessments

The landscape of mandatory digital risk assessments and due diligence now includes diverse jurisdictions such as Australia, Brazil, the EU, India, and the UK among many others. In light of the varying level of commitment to the rule of law and to International Human Rights Law that may exist between these jurisdictions, these may at times reinforce or diverge from risk assessment frameworks promoted by multilateral organizations, investors, and other stakeholders. By contrast with these voluntary frameworks, mandatory risk assessment can afford governments both formal and informal leverage over companies to moderate content, track users, or restrict rights, and vis-a-vis civil society who depend on governments for access to data, funding, or other forms of support. To avoid the potential for risk mitigation and regulatory compliance to become tools for state overreach, it is important to find ways to insulate risk management from political interference.

Key Learnings	Recommendations
Governments may use a variety of informal and formal mechanisms to incentivise platforms to remove lawful content.	Governments and courts should ensure that mandatory risk assessment and related regulations are enforced transparently, with appropriate oversight, consistent with rule of law norms and human rights principles. They should narrow and clarify the scope of risk assessment obligations to ensure they are not understood as obligations to remove lawful content.
Mandatory risk assessments can be used to encourage or justify removal of lawful content and suppress legitimate social and political discourse.	Companies and civil society organisations should advocate for clear, narrowly defined, and transparent risk definitions to prevent political abuse of safety laws. Companies should exercise transparency about compliance with mandatory risk assessment requirements and related impacts on human rights.
Government interventions tend to remain hidden from public scrutiny.	Companies should use public transparency reports, as well as mechanisms such as GNI, to share details of government demands and their responses, where possible. Civil society organisations should use this information to hold governments to account.

3 The Interconnected Rights and Risks Landscape

3.1 Image abuse: its scale, violence and disproportionate impact on women, girls, and marginalized groups

Non-consensual intimate imagery has escalated dramatically in scale and impact, fueled by generative AI. This image abuse can be traumatic for those targeted, silencing participation of women and marginalized groups in digital spaces. Mitigating these harms is complicated by globally varying cultural constructions of intimacy. Reporting and removal must be complemented by work to tackle this abuse in technology and product design. New content policy approaches may also be required.

Key Learnings	Recommendations
Image abuse harms its targets in multiple ways. It can cause them to speak less freely and leave digital spaces. Among other things, image abuse can impact privacy, security, freedom of expression, access to information, and at times the ability to safely study or work.	All actors should seek to create responses and mechanisms that put power into the hands of those who are targeted by image abuse. Companies could do more to raise awareness for users about their rules prohibiting NCII; onboarding education for users about the rules can help proactively reduce the amount of NCII content that's posted. Companies should prioritise rapid, cross-platform content reporting and takedown processes for known image abuse. For example, initiatives that offer users the ability to share images they'd like removed using privacy-preserving device-side hashing technology. These processes should have appropriate rights-based safeguards, so they are not themselves abused. Governments should establish clear, rights-based legal frameworks defining and tackling NCII.

Generative AI is multiplying the scale and impact of image abuse.	Companies and developers should devise and deploy safeguards at product and technology design stages, potentially based on content provenance and varying approaches to consent. These tools should also have appropriate rights-based safeguards, so adult and other legal, non-abusive content is not inadvertently classified as NCII and removed.
Helplines and other response mechanisms provide critical support, but these mechanisms are not able to serve even a fraction of the estimated impacted targets.	Companies, governments, and funders should increase support to helplines and other support for those targeted.
There are cultural differences, contestation, and disagreements about key definitions in this space, including what "intimate" and "consent" means in the online context.	Researchers, civil society organisations, and governments can seek to more clearly define language, and balance power appropriately across sectors, when proposing regulatory frameworks or developing mitigations. For example, there was concern about enabling regulatory definitions of "intimate" in more authoritarian or democratic backsliding contexts. Whereas in other contexts, there was concern about under-balancing the privacy and freedom of expression impacts from NCII. All actors should carefully consider the freedom of expression implications of seeking up-front consent frameworks in the context of stopping NCII.
The creation and spread of abusive images is being monetised and amplified across platforms.	Companies should ban the monetization, advertising, and algorithmic amplification of nudification apps and other tools used to generate abusive images.

3.2 Child safety: rights-based approaches in theory and practice; the variable efficacy and inherent risks of assurance systems

Child sexual abuse and exploitation is a shared global harm requiring action. Unfortunately, the online world was not designed with children in mind. Now, there is a shared responsibility to ensure children can appropriately participate in rights-respecting, safe ways. This burden should not sit with children or be the sole responsibility of parents and caregivers.

However, vague risk assessments and bans increasingly threaten the rights of children and young people in the name of safety. Regulatory frameworks must recognise children and young people as active rights-holders and incentivise the development of online environments that are both safe and rights-respecting. Meaningful protection requires implementing design frameworks that deliver high quality, safe, and right-respecting services.

While age verification or assurance is increasingly mandated, its efficacy is variable and it can generate risks to privacy and security of both children and adults. Internal testing by age assurance technology providers can focus on the wrong metrics and often lacks rigorous external validation. Independent validation, reporting, and formal accreditation mechanisms are needed to ensure the fairness and reliability of age assurance systems.

Key Learnings	Recommendations
It is inappropriate to solely place the burden of online safety on either children or their parents. Youth voices and needs are frequently excluded from safety policy and product design.	Companies should implement policies and frameworks that ensure that both safety and child rights are fully considered and addressed at technology and product design stages. Consideration should be given to mechanisms like youth co-design and regulatory sandboxes.
Child safety measures are too often based on false trade-offs between children's privacy and access to information.	Governments and companies should insist upon rights-based approaches that simultaneously ensure safety and privacy for children, as well as adults.

Broad, overarching social media bans restrict fundamental youth rights and fail to address the root causes of safety issues.	Companies must ensure that products and services are safe by design and default by including appropriate, targeted and effective safety features.
The efficacy, security, privacy, and fairness of age assurance technologies is highly variable and some implementations pose significant privacy concerns.	Companies and technology providers should publish clear and transparent reports detailing how evaluations are conducted and what they show. Governments should take care not to promote the implementation of ineffective or inappropriate age assurance solutions.
Internal testing by age assurance providers is inconsistent and can answer the wrong regulatory questions. A lack of independent verification allows flawed systems to operate and create unwarranted access restrictions.	Governments, civil society organisations, and standards bodies should establish well-defined, consensus-driven standards, such as ISO/IEC 27566-1, to facilitate transparent and accountable age assurance systems.

3.3 Monetisation: motivating illegal content and discriminatory behaviours

Platform monetization features, such as creator funds, affiliate tools, and tip jars are valuable innovations that enable creators to participate, engage, and crowdfund online. However, they also can incentivise the creation and spread of harmful and sometimes illegal content. Greater external scrutiny of monetisation processes may allow companies to better identify related risks and develop stronger mitigation strategies.

Key Learnings	Recommendations
Monetisation features can incentivize the creation of high-engagement, harmful, and illegal content.	Companies should carefully assess the risks associated with the financial incentives, payments, and restrictions that they operationalise.
Services offered by platforms that enable users to make money, and the changes made to these services over time, are not always clearly disclosed.	Companies should be transparent about all monetisation services, rule adjustments, and policy updates. For some companies, monetisation policies sit within different resource hubs than other types of policies, which can be confusing to users.
Platforms can be abused for the transfer of illegal funds.	Companies and regulators should continue to follow or adopt anti-money laundering and financial sector best practices to prevent financial abuse on digital platforms and services.
Monetisation rules can lead to economic hardship and censorship of creators.	Companies should ensure fair terms for creators, robust due diligence, and transparent remedy processes for creators facing demonetisation or other financial restrictions, while adhering to relevant regulations such as sanctions.

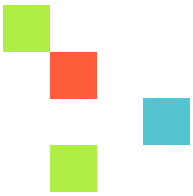
3.4 Artificial Intelligence (AI): accountability gaps

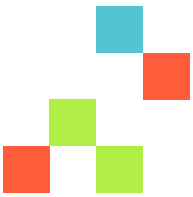
Artificial Intelligence (AI) is rapidly impacting the broader risk environment, user behavior, provider affordances, and internal trust and safety tooling. To date, trust and safety efforts remain disproportionately focused on using AI for content detection and evaluation. The use of AI based detection tools creates risks of overly broad moderation, especially where minority languages and contextual or cultural nuance are involved, therefore creating significant

concerns related to freedom of expression and to the right to non-discrimination. Where content moderation decisions are automated, the governance model should be designed, prior to the deployment of technology, to ensure that there is clear accountability and ultimately human oversight for decision-making. The best way to identify and mitigate risks associated with AI is to subject the development, deployment, or use of AI to human rights-based analysis.

Key Learnings	Recommendations
A disproportionate focus on content detection may lead to governance and accountability being neglected in technology and service designs.	Companies and regulators should focus upstream on AI model development and deployment, insisting on sound governance and accountability processes as well as rigorous safety testing and provenance standards. Companies should ensure appropriate human review of the automation of any aspects of trust and safety or content moderation.
Reliance on automated detection increases the risk of overly broad moderation, especially where minority languages or contextual/ cultural nuance are involved, creating significant concerns related to freedom of expression and to the right to non-discrimination.	Companies should invest equitably in safety support, training data, and moderation resources for minority languages, and exercise heightened due diligence in conflict and high risk contexts.
Platform restrictions on researcher access to data reduce the possibilities for independent analytical oversight of AI tools/use.	Companies and governments should protect and support independent academic and civil society research and analysis of automated decision-making.

Autonomous systems can lack adequate governance frameworks and reconstructable accountability trails.	Companies should ensure and demonstrate reconstructable decision trails in their automated systems, together with effective human oversight structures.
Vague risk definitions push developers and deployers of automated systems toward excessive caution, chilling expression by default.	Companies and developers should avoid broad content filters and hardcoded refusals that create de facto speech restrictions.
Generative AI has no established metrics to detect when content moderation goes too far.	Companies should develop measurable indicators to identify when risk mitigations are themselves inappropriate or disproportionate.
Freedom of expression and privacy are not sufficiently embedded into the practices of developing and deploying AI.	Embed freedom of expression standards directly into AI model and product development and deployment pipelines.





4 Risk Management and Regulatory Practice

4.1 Operationalising risk management in technology companies

Most companies adopt a three-line defense model for risk management, which separates operational management, risk management and compliance, and internal audit. It is challenging to align this with agile product cycles and overly rigid regulations may lead companies to focus more on satisfying regulatory requirements than genuinely ensuring the safety and freedom of expression of users.

Key Learnings	Recommendations
A compliance focus might reduce crucial trust and safety efforts to performative checkbox exercises for regulators. Having processes in place does not guarantee safer, more rights-respecting outcomes.	Companies and regulators should seek evidence of improved outcomes, both in terms of safety and rights, for users and platforms environments. Companies and regulators should seek to share the relevant data with researchers and civil society organisations, so they can conduct evidence-based research on outcomes. Researchers and civil society organisations can continue to develop methodologies and conduct research to generate evidence bases about the efficacy of regulatory frameworks and company implementation efforts.

Consistency of risk assessments across global company operations is challenging to achieve, especially given competing demands and sometimes prescriptive requirements, of regulators.	Companies should adopt standardised, cross-jurisdictional risk taxonomies and establish common vocabularies and formats for describing and evaluating risks. They should share transparently, where they can, about what mechanisms are effective to reduce risks and respect rights.
Companies use their own criteria to determine thresholds for proactive safety and human rights reviews, because of the "critical impact" the changes may have on users and society.	Regulators should define critical impact thresholds more clearly and companies should implement safety and human rights reviews prior to all major product deployments.
Risk management in AI, as with other digital technologies, may involve political and normative choices, making it vulnerable to government overreach.	Define systemic risk tightly to prevent regulatory scope creep and political interference in AI outputs.
Users are largely excluded from the risk governance processes that shape the tools they use.	Bring users into risk management processes as active participants, not just subjects of policy.

4.2 Counter terrorism and risk management

The application of counterterrorism laws to digital services is problematic as moderation tools are not able to sufficiently establish intent and context. Counter terrorism laws should be narrowed before being transposed to digital environments and companies should ensure that human reviewers assess offensive elements to differentiate incitement from legitimate discourse like satire.

Key Learnings	Recommendations
Counterterrorism laws built for physical conduct are poorly suited to online speech, and broad definitions risk criminalizing legitimate digital content.	Governments and companies should distinguish between physical conduct and online speech, and narrow definitions of terrorism-related offences before applying them to digital platforms.
Criminal liability requires consideration of intent and state of mind, which are difficult to establish in online content moderation.	Companies should assess whether all elements of an offence, including mental elements, are reasonably satisfied before classifying content as illegal.
Whether content constitutes an offence is highly context-dependent, requiring consideration of authorship, purpose, and surrounding circumstances.	Companies should design moderation systems to assess whether content constitutes reportage, satire, or commentary rather than incitement.

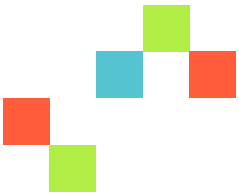
Algorithmic tools may lack the cultural sensitivity to make nuanced judgments at scale, frequently suppressing legitimate political and social discourse.	Companies should supplement algorithmic tools with human review for contextually complex content, particularly satire, political commentary, and content involving designated groups.

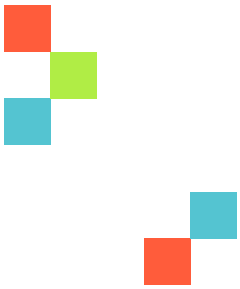
4.3 Gathering civil society insights and evaluating safety by design

Risk assessment is less effective when companies obscure their methodologies and researchers do not have access to data. For current regulatory approaches to function well, strong internal governance, internal due diligence, and independent verification must be applied to the entirety of business operations.

Key Learnings	Recommendations
Companies are not sufficiently transparent about their risk assessment methodologies or the effectiveness of their mitigations.	Companies should be more open about how their risk assessments are conducted, informed, reviewed and verified.
Independent researchers do not have sufficient access to data to be able to properly evaluate company risk assessments.	Governments and companies should ensure that researchers receive timely, comprehensive, and usable data, and should otherwise work to support credible, independent research around digital risks and mitigations.

Whether content constitutes an offence is highly context-dependent, requiring consideration of authorship, purpose, and surrounding circumstances.	Companies should design moderation systems to assess whether content constitutes reportage, satire, or commentary rather than incitement.
Risk assessments are sometimes conducted in a vacuum, ignoring the impacts of the broader corporate business model.	Companies should link operational risk assessments together and demonstrate their understanding of the societal impact of their entire operation.
Users are treated as passive risk subjects rather than autonomous individuals with diverse, legitimate needs.	Companies, civil society should design and advocate for systems that respect individual user agency and account for diverse use cases.





Conclusion

The 2026 Rights & Risks Forum enabled a range of academic, civil society, and private sector experts to discuss how to keep rights-based approaches at the center of conceptualizing and assessing risks associated with digital platforms and services.

DTSP and GNI look forward to continuing these conversations in other fora and through this annual event series.

Annex: Reading list

These are some of the civil society and academic resources that the organisers used to prepare for the Forum, that participants shared after the event, and that speakers mentioned during their remarks. This is not a comprehensive list; it is intended to provide a jumping off point that might be useful to the various stakeholders in this space who are conducting, considering, and evaluating risk assessments from a human rights perspective. This primarily focuses on resources that have come out in the last year. Please note that inclusion is not an endorsement of positions or viewpoints.

On risks, assessment and mitigation, rights frameworks, and politicisation:

- Owen Bennett: Why Europe Could Block X Over Grok Scandal But Probably Won't | TechPolicy.Press
- Tim Bernard: Platforms Report to EU Regulators Under DSA With an Eye on US Politics
- Tim Bernard: Reading the Systemic Risk Assessments for Major Speech Platforms: Notes and Observations
- Mateus Correia de Carvalho and Rachel Griffin: Civil society participation and participatory justice in DSA systemic risk management
- CELE: Are Risks the New Rights? The Perils of Risk-based Approaches to Speech Regulation
- Council of Europe: Recommendation of the Committee of Ministers to member States on online safety and empowerment of users and content creators
- DSA Civil Society Coordination Group: Initial Analysis on the First Round of Risk Assessments Reports under the EU Digital Services Act

- DSA Human Rights Alliance: Principles Calling for DSA Enforcement to Incorporate Global Perspectives
- European Partnership for Democracy & Liberties: The Digital Services Act: Weak Democratic Safeguards on Big Tech
- Future of Free Speech: Europe Cannot Protect Democracy by Distrusting Its Citizens
- Jacques Delors Centre: How Has the DSA Performed in Protecting Election Integrity? | TechPolicy.Press
- Daphne Keller: The Rise of the Compliant Speech Platform
- Kate Klonick: The State Department's X Directive and the End of Platform Independence | Lawfare
- Knight-Georgetown Institute: Measuring Risk: What EU Risk Assessments and US Litigation Reveal About Meta and TikTok
- Prosocial Tech Design: Regulation: A Practical Guide
- Afsaneh Rigot: Design from the Margins
- SMEX: Meta's 2024 Human Rights Report: too little, too late
- SMEX: Confronting Structural Silencing: Challenges and Resistance among Feminist Activists in Lebanon
- Tech Policy Press: The Digital Services Act is a Lightning Rod for Debate
- WHAT TO FIX: De-Risking Social Media Monetisation
- WHAT TO FIX: How Facebook (Continues to) Channel Money to Sanctioned Entities

On AI as a risk and a mitigation tool:

- BSR: Human Rights Assessment of the genAI Value Chain & Responsible AI Practitioner Guides
- Columbia University's School of International and Public Affairs' Institute for Global Politics: "AI Slop" and the Information Ecosystem: Insights from a Cross-Sector Convening
- Future of Free Speech: Artificial Intelligence and Freedom of Expression in the European Union
- Future of Free Speech: That Violates My Policies: AI Laws, Chatbots, and the Future of Expression
- OECD: Due Diligence Guidance for Responsible AI
- Oversight Board: Content Moderation in a New Era for AI and Automation
- Wikimedia Foundation & Taraaz Research: Artificial Intelligence and Machine Learning Human Rights Impact Assessment

On child safety risks:

- Center for Democracy & Technology: What Kids and Parents Want: Policy Insights for Social Media Safety Features
- Children's Online Redress Sandbox: COR Sandbox

- Tech Coalition: Voluntary Framework for Industry Transparency
- Tech Legality: Online Platform Regulation and Children's Rights and Safety in a Digital World: A Global Comparative Analysis
- Knight-Georgetown Institute: Age Assurance Online: A Technical Assessment of Current Systems and their Limitations
- Tijana Milosevic, Anne Collier, Elisabeth Staksrud, and Ioanna Noula: Exploring the role of Dignity in Design for Artificial Intelligence-Based Cyberbullying Interventions
- WeProtect Global Alliance: Global Threat Assessment 2025 Preventing technology- facilitated child sexual exploitation and abuse: From insights to action
- UNICEF: Certification Schemes
- UNICEF: Impact and Risk Assessments
- WeProtect Global Alliance: Prevention Framework
- World Economic Forum: Global Principles on Digital Safety: Translating International Human Rights for the Digital Context

On non-consensual intimate imagery (NCII) risks:

- Council on Tech and Social Cohesion, Integrity Institute & Search for Common Ground: Prevention by Design; Tackling TFGBV at the Source
- Search for Common Ground: Curriculum for Digital Community Stewards on TFGBV
- SMEX: 80% of Women in Lebanon Face Digital Violence
- SWGFL: Closing the Gaps: How Collective Action Can Make NCII Prevention Work
- SWGFL: The Scale of Non-Consensual Intimate Image (NCII) Abuse: A Data-Driven Global Analysis
- WITNESS: The Grok Disaster Isn't An Anomaly. It Follows Warnings That Were Ignored
- WITNESS: Public Comment to Meta Oversight Board on AI-Generated Sexual Exploitation

GLOBAL NETWORK INITIATIVE

Digital Trust & Safety Partnership