Assessing and Mitigating Conflict-Related Online Risks

Challenges for Governments, Regulators and Online Platforms

Helena Schwertheim, Sophie Scheuble & Carolin von Bredow

About the Digital Policy Lab

The Digital Policy Lab (DPL) is an inter-governmental working group focused on charting the policy path forward to prevent and counter the spread of information operations, hate speech, extremist and terrorist content online. It is convened by the Institute for Strategic Dialogue (ISD) and composed of representatives of relevant ministries and regulatory bodies from selected liberal democracies. The DPL aims to foster inter-governmental exchange, provide policymakers and regulators with access to sector-leading expertise and research, and build an international community of practice around key challenges in the digital policy space.

About this paper

As part of the DPL, ISD organised two working group meetings on the topic of assessing and mitigating conflict-related online risks in April 2025. The working group consisted of DPL members including regulators, competent authorities and law enforcement representatives from multiple jurisdictions, as well as representatives from civil society and academia. While participants' contributions have informed the analysis in this paper, the views expressed within do not necessarily reflect the views of all participants, nor any governments involved in this project.

Acknowledgements

We would like to thank all members and participants of the working group for their contributions. In particular, we would like to thank speakers in the two sessions: Aleksandra Atanasova (Reset Tech), Felix Kröner (formerly Reset Tech), Marwa Fatafta (Access Now), Richard Kuchta (ISD Germany), and Friedhelm Weinberg (Mnemonic). Additionally, we would like to thank participants from the following ministries and regulatory authorities: The eSafety Commissioner (Australia), Canadian Heritage (Canada), Public Safety Canada (Canada), Autorità per la Garanzie nelle Comunicazioni (Italy), Classification Office (New Zealand), Ofcom (United Kingdom). We also thank Article 19, Centre for Information Resilience, Centre for the Study of Organized Hate, Digihub Africa, and the Forum on Information & Democracy for their participation in the working group.

About the authors

Helena Schwertheim is a Senior Digital Policy and Research Manager at ISD. She leads the Digital Policy Lab (DPL), an intergovernmental working group focused on policy responses to prevent and counter online harms. As part of the Digital Policy Team, Helena advises key governments, international organisations and tech companies, and collaborates with ISD's Digital Analysis Unit to translate research into actionable digital policy recommendations, with a focus on Technology Facilitated Gender-Based Violence (TFGBV). Previously, Helena managed digital policy and research projects at Democracy Reporting International. She also has experience working in risk and political analysis in international organisations and think tanks including at the UN World Food Programme in Rome and the think tank International IDEA in Stockholm.

Sophie Scheuble is a Digital Policy Coordinator at ISD. She researches the intersection of online safety and extremism across ISD's digital policy portfolio. As part of the Digital Policy Lab (DPL), Sophie coordinates the working groups, annual Summits and contributes to the thematic policy briefs. She previously served as Deputy Director of the International Department at Violence Prevention Network and worked for the European Commission's Radicalisation Awareness Network (RAN) Practitioners on multistakeholder approaches to preventing and countering violent extremism (P/CVE).

Carolin von Bredow is a Digital Policy Associate at ISD, where she contributes to ISD's digital policy advisory work, including the Digital Policy Lab (DPL) and research and engagement on key digital regulation. Carolin also contributes research to ISD's growing technology-facilitated gender-based violence (TFGBV) portfolio and work on artificial intelligence. Prior to ISD, Carolin worked in the Data and Digital Sector at the European Union Agency for Fundamental Rights (FRA) and managed research projects at the Technical University of Munich. Carolin holds a MSc in Politics and Technology from the Technical University of Munich, and a BA in International Relations and Politics from the University of East Anglia.

ALFRED LANDECKER FOUNDATION

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Contents

Executive Summary

As over 120 conflicts play out across the globe, the information environment has become a strategic weapon to rally support, spread propaganda, incite violence and undermine trust.[1] Conflict-related online risks increasingly transcend borders, affecting not only those in conflict zones but also diaspora or minority communities elsewhere, as well as political discourse and social cohesion in countries far from the fighting. These can include online risks such as disinformation and foreign information manipulation and interference (FIMI), terrorist and violent extremist content (TVEC), illegal hate speech[2] and incitement to violence, gender-based violence (GBV) – including when technology facilitated (TFGBV) – and the algorithmic amplification and monetisation of these risks. Over-zealous or inaccurate moderation or suppression of conflict-related content by online platforms can also create significant risks to legitimate and protected freedom of expression.

While platform policies and enforcement often fail in conflict-affected regions where domestic governance and regulatory oversight can be weak, new policy frameworks in liberal democracies such as the European Union's (EU) Digital Services Act (DSA) and the United Kingdom's (UK) Online Safety Act (OSA) offer a crucial opportunity for improved responses and greater accountability. These regulations can protect domestic users affected by online risks emanating from conflict zones, stem the spread of illegal content, provide greater transparency, and help to develop and define best practices for platform responses. They provide tools to safeguard information integrity and online safety such as systemic risk assessments, crisis protocols, and transparency provisions. Strengthening and utilising these mechanisms can help ensure platforms identify and mitigate conflict-related risks, improve rapid responses to emerging crises, and enhance transparency and data access.

This policy brief examines how governments, regulators and technology companies can better identify, mitigate and respond to these online risks. It maps the spectrum of conflict-related risks – ranging from disinformation and hate speech to the over-removal of legitimate content – and considers the role of platform design and market incentives in enabling or exacerbating these risks. It then provides an analysis of the international legal landscape and relevant non-regulatory initiatives, and how current legislation in the EU and the UK can address such risks.

By comparing these regulatory frameworks, the brief identifies potential ways in which they can contribute to mitigating conflict-related online risks. It also highlights the need for stronger cross-border coordination and greater transparency. Mitigating online risks from foreign conflicts is not the primary aim of these regulatory regimes but the tools they provide (if applied with a conflict-sensitive lens) can strengthen information integrity and limit spillover risks.

The brief finally provides recommendations for governments, regulators, platforms and international organisations to safeguard fundamental rights domestically and abroad, and set international benchmarks for platform accountability in conflict settings.

For governments and regulators

- **Apply existing risk frameworks to conflict-related risks:** Governments, regulators and platforms should ensure that the systemic risk categories already established under the DSA (including illegal content, fundamental rights, civic discourse and electoral processes, public security, gender-based violence, and the protection of public health and minors' well-being) are applied to conflict-related harms. This includes those impacting diaspora communities and humanitarian actors. In the UK, existing duties around illegal content, children's safety, and the Guidance on safer life online for women and girls should likewise be interpreted and enforced to address harms linked to conflicts.

- **Facilitate structured cross-border cooperation:** Establish formal, ongoing channels for regulators, governments, and civil society organisations to exchange information on conflict-related online risks, both during and after crises. This could include:

    - **Sharing case studies and lessons learned** when powers such as the DSA's crisis response mechanism (Article 36) or systemic risk assessment obligations (Articles 34–35), or the UK OSA's illegal content and children's safety duties, have been applied in a conflict-related context.

- **Developing joint rapid-response protocols** to ensure consistent action across jurisdictions when harmful content – such as incitement to violence, terrorist material, or disinformation – spills over borders.

- **Including civil society as an early-warning system**, enabling regulators to act quickly on local intelligence about emerging risks.

- **Using international coordination forums** (e.g., European Board for Digital Services, the Global Online Safety Regulators Network (GOSRN), or other multilateral online safety networks) to align approaches and avoid contradictory or duplicative interventions.

- **Recognise and address the domestic impact of conflict-related online risks:** Acknowledge that online risks originating in foreign conflicts, such as illegal hate speech, terrorist content, disinformation and FIMI, and TFGBV can spill over into domestic information environments, fuelling polarisation, hate speech and offline tensions. Regulators should assess whether conflicts may have an outsized impact in their jurisdictions. They can then make use of existing emergency and systemic risk management powers (such as the EU DSA's Article 36 crisis response mechanism and the UK OSA's Codes of Practice) to require platforms to take timely, proportionate, and rights-respecting measures on conflict-related content. This must be transparent when utilised, and should include activating crisis protocols, scaling moderation capacity (including in relevant languages), and engaging local and diaspora civil society to ensure context-specific responses.

- **Safeguard fundamental rights and preserve evidence in crisis response:** National frameworks such as the UK OSA and its codes, and crisis protocols under the EU DSA (Article 36) should require platforms to securely retain content with potential evidentiary value following removal (including material relevant to international investigations, human rights documentation or humanitarian protection). Clear protocols should ensure this is done in compliance with local laws, international human rights standards, with mechanisms for timely, secure sharing with competent authorities, trusted civil society partners and verified investigators. These measures should be accompanied by transparent reporting on takedown decisions and rapid redress mechanisms to prevent and reverse wrongful removals. This ensures that emergency actions remain proportionate, rights-respecting and subject to independent oversight. Lessons from past conflicts, where automated takedowns erased potential war crimes evidence, underscore the need for such requirements. These should be developed in consultation with conflict-affected communities, humanitarian and human rights experts, and local civil society.

For platforms

- **Conduct proactive, localised risk assessments in fragile and conflict-affected contexts and adjust mitigations proportionately**: In line with the UN Guiding Principles on Business and Human Rights (UNGP), platforms should identify, assess and address actual and potential human rights impacts, including conflict-related risks; they should publish these assessments for public scrutiny. An illustration of why assessments must be timely, forward-looking and methodologically robust is Facebook's (now Meta) Human Rights Impact Assessment (HRIA) in Myanmar, commissioned in 2018 after atrocities had already unfolded.[3] Where applicable, platforms must also meet their legal duties under the EU DSA (Articles 34–35) to assess systemic risks, ensuring assessments explicitly include cross-border harms and impacts on diaspora communities. Rather than relying on reactive crisis mitigations, platforms should anticipate surges in harmful content during crisis moments and adapt moderation capacity, language coverage and escalation processes with adequate foresight.

- **Develop and test rapid-response crisis protocols**: Establish and rehearse procedures to respond swiftly to conflict-related online risks, aligned with DSA Article 36 crisis response provisions and relevant national safety codes. Integrate these protocols with government and civil society early-warning mechanisms to ensure rapid threat mitigation while preserving content that may have evidentiary value for international investigations or humanitarian purposes.

- **Coordinate across industry to address cross-platform risks**: Establish and strengthen cross-platform cooperation to address conflict-related risks. This could include removing links, mirrors and re-uploads of terrorist, violent extremist or hate-inciting material. It has been demonstrated how pro-Kremlin propaganda networks adapted after the EU's March 2022 sanctions by using coordinated account networks across platforms.[4] Platforms should learn

and build on existing cross-platform mechanisms (such as the Christchurch Call's Crisis Response Protocol and the Global Internet Forum to Counter Terrorism's (GIFCT) Incident Response Framework for terrorist and violent extremist content) and adopt similar models explicitly for conflict-related risks.

- **Improve language and contextual expertise**: In line with the UNGP's call for "meaningful stakeholder engagement", platforms should enhance detection of harmful content through meaningful local-language engagement, particularly in under-represented and conflict-affected regions. Experiences in Ethiopia illustrate the stakes: Facebook (now Meta) added Amharic and Oromo moderation only after ethnic violence had escalate. The platform also heavily relied on AI tools poorly trained for Amharic or Tigrinya, resulting in delayed action against incitement and hate speech.[5] Platforms should proactively partner with trusted local actors to close these enforcement gaps, ensuring platform policies are enforced equitably across linguistic contexts.

For multilateral and international organisations

- **Create model online crisis response protocols beyond Terrorist and Violent Extremist Content (TVEC)**: Develop adaptable protocols for responding to conflict-related risks such as disinformation, hate speech, targeted harassment, and other systemic risks during conflict. This should draw on lessons from TVEC frameworks such as the Christchurch Call's Crisis Response Protocol and the GIFCT Content Incident Protocol, which already coordinate real-time, cross-platform action for TVEC. These protocols should embed safeguards for fundamental rights including activation criteria, evidence preservation, proportionate enforcement and rapid redress mechanisms (in line with international human rights law).

- **Strengthen multi-stakeholder early-warning systems by connecting existing humanitarian and research networks**: Existing initiatives – such as the UN Refugee Agency (UNHCR) Information Integrity Toolkit,[6] the Risk Communication and Community Engagement (RCCE) Collective Service,[7] the European Digital Media Observatory (EDMO) Hubs,[8] and the EU's Rapid Alert System[9] – should be connected to civil society reporting networks and platform risk monitoring teams to detect and flag emerging online threats in conflict-affected and at-risk communities. This standardisation into a common alert schema could be based in the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) Data Responsibility Framework; this would ensure standardised and safe information sharing to and from existing humanitarian "rumour-tracking" efforts.[10] These systems should also prioritise inclusivity, bringing in perspectives from affected regions and diaspora communities to improve contextual accuracy and responsiveness.

- **Convene and facilitate cross-platform crisis coordination that also covers small and fringe services for online conflict risks beyond TVEC**: Neutral conveners (such as UN OCHA's Humanitarian Cluster System, or regional entities such as EDMO) should align and synchronise actions between global, regional, and smaller or fringe platforms during conflict-related crises. This ensures a consistent, time-bound, rights-respecting approach. Facilitation should include secure information-sharing channels, technical assistance for smaller platforms lacking capacity, and joint monitoring mechanisms to track the effectiveness and proportionality of interventions.

Introduction: the role of the information environment in conflict

In 2024, there were more than 120 active conflicts across the globe (61 state-based and 74 non-state) according to the Peace Research Institute Oslo (PRIO).[11] Many of these conflicts remain outside the spotlight. Underreported crises slip from international attention, enabling violations of international law and human rights to go largely unnoticed. Mapping conflict-related online risks requires bridging insights from conflict analysis, humanitarian protection and online safety research.

In today's conflicts, digital spaces have become a central arena of contestation. Online platforms shape how wars are conducted and perceived, providing channels for influence operations, recruitment, and mobilisation by both state and non-state actors. At the same time, these spaces carry risks for affected populations, diaspora communities, and international audiences, as hostile narratives and harmful content spill across borders. Understanding the role of the online information environment is therefore essential to assessing and mitigating conflict-related risks.

Scope of the problem and types of online risks

A variety of typologies and frameworks already guide the identification and analysis of risks in conflict settings and provide a valuable tool for understanding their impact abroad. The International Committee of the Red Cross (ICRC) has developed a response framework for humanitarian organisations addressing harmful information in conflict settings.[12] Within the framework, harmful information refers to content that causes or contributes to physical, psychological, economic or social harm. The ICRC categorises this into four core types:

- Misinformation (false information shared without malicious intent),
- Disinformation (deliberately false information disseminated intentionally),
- Malinformation (factual information shared with malicious intent),
- Hate speech.

The framework also highlights other forms of harmful narratives that may violate international humanitarian law (IHL). This includes content that dehumanises adversaries, incites violence, undermines legal norms or erodes trust in humanitarian action.

While the ICRC framework focuses on the humanitarian implications of harmful information, these risks intersect with a wider range of online risks that can influence conflict dynamics and spill over into international information environments. These can include:

- **Hate speech and incitement to violence**: The United Nation (UN) Strategy and Plan of Action on Hate Speech defines hate speech as "any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of their identity".[13] In armed conflicts, such narratives target ethnic, religious or political groups, women, humanitarian actors or journalists. They can potentially fuel offline violence both in conflict-affected states and abroad.

- **Terrorist content:** Terrorist actors have exploited online platforms to disseminate battlefield media, propaganda, martyrdom videos, and fundraising appeals. UN Security Council Resolutions 1624 (2005)[14] and 2354 (2017)[15] established international obligations to prohibit and prevent incitement to terrorism. The EU Terrorist Content Online Regulation (2021/784)[16] defines terrorist content as material that incites or glorifies terrorist acts, provides instructional material, or promotes or solicits support for proscribed terrorist groups. Such content is often prominent in the broader information environment relating to armed conflicts such as those in Syria, Somalia, the Sahel, Ukraine, Afghanistan and Israel/Hamas. Because some groups are both conflict parties and terrorist-listed actors, there is a heightened risk of the over-removal of evidence relating to war crimes and other content.

- **Disinformation**: While the UN General Assembly has not adopted an overarching binding definition of disinformation, the resolution on countering disinformation for the promotion and protection of human rights and fundamental freedoms (A/RES/76/227) frames it as the deliberate dissemination of "false or misleading information with the potential to cause public harm".[17] In conflict contexts, disinformation campaigns aim to distort

public opinion of the conflict and its parties, undermine legal norms, and inflame tensions. Most descriptions reference the (coordinated) spread of false or misleading narratives to shape public opinion of conflicts and their parties, inflame tensions, or undermine democratic processes, often amplified through AI-generated content and cross-platform campaigns.

- **Foreign information manipulation and interference (FIMI)**: FIMI refers to coordinated and deceptive behaviours that seek to influence audiences, often without regard to the veracity of the content itself. The European External Action Service (EEAS) defines FIMI as "the coordinated and intentional manipulation of the information environment by foreign actors. Such activity is manipulative, deceptive and often conducted by foreign actors."[18] This can include cross-platform amplification, use of fake accounts and AI-assisted assets.

- **Technology-facilitated gender-based violence (TFGBV):** The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) General Recommendation No. 35 (2017)[19] and the UN Special Rapporteur on violence against women recognises that gender-based violence also occurs through technology-mediated environments. The ICRC (2024) defines TFGBV in the humanitarian and conflict context as "any act of gender-based violence against women that is committed, assisted or aggravated... by the use of information communication technologies (ICT)."[20]

- **Over-moderation and content suppression:** The premature or erroneous removal of lawful content, such as evidence of human rights abuses, media reporting or activist content can hinder accountability efforts and skew public understanding of events. It can also violate the freedom of expression of affected communities in and beyond the zone of conflict. The UN Human Rights Committee, General Comment No. 34 (2011) underscores that restrictions on expression must meet strict tests of legality, necessity and proportionality.

- **Algorithmic amplification and monetisation:** Research indicates that engagement-driven platform design can elevate divisive conflict-related content.[21] Engagement-based monetisation of content or donation features can create financial rewards for sensational or misleading conflict narratives.[22] These dynamics can have material consequences in wars: for example, in Myanmar, UN investigators and Amnesty International linked Facebook's recommender systems to the amplification of anti-Rohingya hate and incitement to violence.[23]

These risks can have global impacts including surges in targeted harassment and hate crimes against diaspora communities as well as broader waves of identity-based harms.[24] Conflict-related narratives can deepen divisions and reproduce the conflict dynamics across borders, polarising political debates in distant countries, and triggering offline violence in countries far from the conflict.[25]

Online risks – perpetrated, facilitated, or amplified on social media platforms – often occur in jurisdictions where companies operate with limited scrutiny, oversight or accountability. In conflict settings, platforms' Terms of Service (ToS), community guidelines, crisis protocols and global policies should theoretically safeguard affected communities and mitigate spillover risks abroad. Yet investigations[26] repeatedly[27] show inconsistent or absent enforcement, with some platform systems and design choices actively driving conflict dynamics.

In many conflict zones, internet governance is understandably a lower priority for domestic authorities and enforcement capacity may be entirely absent. This leaves international humanitarian, human rights law, and norms – such as the United Nations Guiding Principles on Business and Human Rights (UNGPs) – as the primary, and often insufficient, basis for holding platforms to account. While civil society actors have developed non-regulatory initiatives[28] to urge platforms to identify and counter conflict-related online risks, their effectiveness depends largely on the goodwill and voluntary cooperation of tech companies. Without binding obligations, tech companies can continue to profit from engagement with harmful content while externalising the societal costs.

Relevance to digital regulation and platform governance

This regulatory vacuum in conflict-affected regions underscores the importance of robust regulatory regimes, such as the EU's DSA, and the UK's OSA. These regulations primarily protect users within their jurisdictions and can help to stem the spread of conflict-related risks, while also potentially spurring improvements in platform responses to online risks by encouraging the introduction of new safety features or approaches that platforms may decide to introduce globally.

Such regulations also contain various provisions that can be leveraged to mitigate conflict-related risks. These

include systemic risk assessments, crisis protocols, and transparency and data access obligations. Effective implementation and enforcement of such regulation may:

- Require platforms to identify and mitigate conflict-related systemic risks, including those arising outside their primary markets,
- Strengthen crisis response protocols to respond rapidly to emerging online threats during conflicts,
- Improve transparency and data access for independent monitoring,
- Encourage meaningful engagement with local and diaspora civil society actors to ensure context-specific moderation.

These regulations primarily protect residents within their jurisdictions, but they can also help to mitigate the impact of illegal content and coordinated FIMI campaigns from conflict zones, reducing risks domestically and safeguarding reliable information essential for decision-making and safety in wartime. At the same time, they set democratic governance precedents and offer reference points for platforms and policymakers, supporting more coherent international approaches.

This policy brief examines how such measures can be applied and strengthened to address the unique challenges of online risks in conflict contexts, drawing on lessons from recent conflicts, existing regulatory frameworks, and international best practice.

Cross-border impacts of conflict-related online risks

This section maps conflict-related online risks whose impacts extend beyond the borders of conflict zones. These affect communities, diaspora groups, social cohesion and democratic discourse in other jurisdictions. Understanding these risks is essential for assessing how existing regulatory frameworks are relevant and can be applied to mitigate their spread and uphold information integrity across borders.

Disinformation and information manipulation

Access to trustworthy information is particularly fragile during conflict, but this challenge extends beyond the warzone. Conflict-related disinformation and foreign information manipulation and influence (FIMI) can quickly shape public perception, mobilise or demoralise communities, and polarise societies in countries thousands of miles from the fighting. Social media platforms and private messaging apps enable this reach; manipulated narratives extend beyond directly affected populations to diaspora communities and international audiences whose perceptions may influence political discourse, international and humanitarian responses. At its worst, the spread of disinformation and FIMI may trigger offline tensions and violence far beyond the immediate conflict zone.

During the Iran-Israel war in June 2025, synthetic media played a prominent role in shaping public perception signalling an escalation in the use of artificial intelligence (AI)-generated content in wartime information operations. One video, widely shared online and amplified by Israeli officials,[29] appeared to show a precise strike on the gates of Evin prison in Tehran but was later found to likely be AI-generated.[30] AI-generated content had been observed in other wars, including in Ukraine and Israel-Hamas. However, the scale, speed, and far-reaching use of generative AI tools during this '12-day war' contributed to a highly distorted information environment. In some cases, authentic imagery of air strikes was falsely dismissed as AI-generated.[31] Similar patterns emerged during the 2025 India–Pakistan confrontation, when deepfakes and misattributed conflict footage spread rapidly across borders, fuelling both domestic tensions and diplomatic friction.[32]

While much of the attention around disinformation and information manipulation focuses on its role in active conflicts, its reach extends far beyond the conflict zones. The ICRC recognises that online risks including disinformation, influence operations, and cyber activities are not confined to the immediate geography of war. The cross-border flow of manipulated information and cyber-enabled targeting increasingly affects humanitarian operations, protected persons and diaspora communities in neighbouring or even distant states.[33] Diaspora communities may also be targeted with tailored narratives to inflame divisions and domestic audiences may be exposed to distorted portrayals that influence political debates.[34] For example, the Eritrean government has mobilised loyalist networks through diaspora-facing information channels around cultural festivals and community spaces to discredit critics and inflame intra-community rifts.[35] This has led to violent clashes between pro-democracy and pro-government diaspora communities in the Netherlands, Canada, Germany, UK and US.[36]

Targeted attacks, hate speech and harassment

In conflicts with global resonance, online hate and harassment rarely remain confined to the conflict zone. Escalations on the ground can trigger sharp increases in targeted abuse against affected communities abroad, particularly minority and/or diaspora populations, journalists, human rights defenders, and activists engaged in advocacy or public debate.

ISD's monitoring of disinformation and hate in the UK, France, and Germany following the outbreak of the Israel–Hamas war illustrates this spillover effect.[37] Online narratives originating from conflict zones were amplified across European social media, spreading antisemitic and anti-Muslim hate often framed to polarise, intimidate and discredit individuals based on their identity or perceived political stance.[38] These campaigns have included doxing, defamation, coordinated harassment, and hate speech aimed at silencing voices and fracturing community cohesion.

Such targeting inflicts harm on individuals and undermines democratic discourse. Current platform reporting systems are poorly equipped to address patterns of coordinated abuse, often treating incidents in isolation and lacking transparency in moderation outcomes. This gap allows cross-border harassment campaigns to persist, while highlighting the transnational nature of conflict-related online hate and harassment.

This spillover of targeted hate and harassment is on a spectrum of identity-based risks linked to conflict dynamics. Online attacks frequently target characteristics such as gender, ethnicity or race, language, religion, disability, age, sexual orientation, or political affiliation.[39] Women human rights defenders (WHRDs) are especially exposed to gendered forms of online violence. Research conducted in Colombia, Ecuador and Peru finds that digital abuse is frequently linked to the affected individuals' identities (e.g., gender, race, ethnicity, or sexuality) rather than just the content of their work.[40] In some instances, attacks have involved family members and support networks; documented effects include self-censorship and reduced participation in online spaces.

As underscored in the 2022 EU–US Joint Statement on Protecting Human Rights Defenders Online, the digital targeting of civic actors is both a human rights issue, and a broader threat to democratic participation and international security.[41] The statement affirms that digital technologies must not be weaponised to silence human rights defenders. It calls on both governments and technology companies to uphold obligations under international human rights law, including through proactive measures to prevent abuse and support victims. Building on this commitment, there is a need to strengthen reporting mechanisms, improve transparency, ensure survivor-centred platform design, and developing risk mitigation strategies that are tailored to conflict and post-conflict settings.

Over-moderation: concerns for censorship and the documentation of war crimes and human rights investigations

Online platforms involved in the dissemination of conflict-related information play a crucial role in human rights investigations both within and beyond conflict zones. While these platforms are urged under international frameworks such as the UNGP to moderate content that may breach international human rights violations – such as the right to freedom from discrimination and the prohibition of incitement to violence or hatred (which are outlined in more detail on page 13) – they must also address terrorist content or material that facilitates war crimes.

However, over-moderation of conflict-related content on social media platforms carries two serious risks. The first is the violation of freedom of expression through the wrongful removal of lawful speech, particularly that of marginalised voices;[42] the second, the erasure of vital evidence needed for international investigations, human rights documentation, and accountability processes.[43]

Over-moderation poses a serious threat to freedom of expression, as highlighted by organisations like Mnemonic.[44] Platforms' content moderation teams' reliance on sometimes vague definitions in Terms of Services, opaque content moderation algorithms and non-standardised keyword databases can result in the wrongful removal of legitimate legal speech. This includes vital human rights documentation, activism and counter-extremism initiatives. Over-moderation may result from under-resourced content moderation teams and related automated systems, lacking local language expertise, as found in research on conflict zones such as Ethiopia and Myanmar.[45] These systems frequently remove content before it can be viewed, lacking transparency regarding error rates and disproportionately silencing marginalised voices.

In other instances, platforms' content moderation policies lack the readiness to handle conflict-related risks, such as in the case of the Israel-Hamas conflict following its escalation since the 7 October 2023 attacks. Human Rights Watch has found more than 1,050 cases of content removal and restrictions on Instagram and Facebook, which it assessed as affecting lawful expressions of support for Palestinians.[46] Access Now has found similar infringements on freedom of expression.[47] The Business & Human Rights Centre received a response from Meta concerning this type of suppression, but other companies including TikTok "did not respond to allegations of contributing to the suppression of Palestinian voices during times of crisis".[48] Such practices undermine democratic principles, hinder accountability for abuses and obstruct public access to critical information regarding conflicts.

In addition to wrongful or excessive takedowns, there is also the problem of correct removals without preservation. Platforms can appropriately remove genuinely violative content (such as terrorist propaganda, incitement to violence or graphic war material) but often fail to securely archive it before deletion. This creates a significant accountability gap: material with potential evidentiary value for war crimes prosecutions, human rights documentation or transitional justice is permanently lost. This has hindered international investigations in Syria, Ethiopia and Myanmar, forcing civil society initiatives like Mnemonic's Syrian Archive to step in, capturing and storing hundreds of thousands of deleted records for use in legal proceedings. While these efforts demonstrate that secure archiving is possible, they remain ad-hoc and underfunded, rather than systematically embedded into platform crisis protocols.

Organisations dedicated to archiving content on human rights violations face challenges posed by content removals. Organisations such as Mnemonic play a crucial role in preserving relevant data by gathering, verifying and investigating digital information that highlights human rights abuses. These efforts serve as a digital record of these violations and provide essential evidence for legal cases. Mnemonic's Syrian Archives' 'Lost and Found' campaign has stored 350,357 videos and 650,000+ records to social media platforms to assist in "criminal case building as well as human rights research"[49]. Without deliberate safeguards for evidence preservation, such irreplaceable records risk being lost to automated takedowns.

This dynamic underscores a central challenge for regulators and platforms alike: the very categories of content to be removed are often those most relevant for documenting human rights abuses and investigating international crimes. Effective crisis response requires a careful balance between the imperative to reduce immediate online risks and the equally vital need to preserve evidence, safeguard freedom of expression and ensure accountability.

Considerations: platform design, vulnerabilities, and market incentives

Platform design choices, content curation or moderation systems and the resulting incentives they create can also play a decisive role. They can enable or amplify conflict-related online risks, whether through deliberate misuse by bad actors or through design choices that reward engagement with harmful content.

Algorithmic and design risks: Recommender systems on major platforms can privilege emotionally charged, polarising content by prioritising content which receives high levels of engagement. Engagement-based optimisation can amplify harmful narratives over constructive or factual ones, including during moments of heightened conflict.[50] This algorithmic prioritisation can sustain harmful narratives for extended periods, even when they are identified as false; this is especially true when it is combined with opaque or inconsistent labelling of manipulated content.[51]

Transparency and representation gaps: Platforms provide limited visibility into how moderation and ranking systems operate, obscuring their role in amplifying harmful content. Research into this dynamic during conflict shows that harmful narratives often circulate for longer in non-English contexts. This reflects the underrepresentation of local expertise in moderation teams and weaker enforcement outside of high-priority markets.[52]

Monetisation and online advertising: Online platforms and generative AI companies continue to profit from engagement, including with harmful conflict-related content. Disinformation and clickbait is monetised across multiple services, with limited effective safeguards to prevent revenue generation from false or harmful narratives. Research by Reset Tech in 2024 during the full-scale invasion of Ukraine examined how social media platforms and advertising systems monetise disinformation and propaganda related to the conflict.[53] The study investigated whether platform ad policies designed to prohibit monetisation of such content were enforced in practice. It found that content from known disinformation sources, including Russian state-linked outlets, was still being monetised via advertising, despite explicit policy bans; in some cases, this content benefited from cross-platform monetisation opportunities. These findings illustrate how existing market incentives, combined with weak or inconsistent enforcement of ad policies, can directly undermine conflict-related harm mitigation efforts.

Governance and coordination deficits: Harmful narratives frequently migrate between platforms, exploiting differences in enforcement. Reset Tech documents "cross-platform monetisation" strategies in which disinformation actors can adapt content and tactics to bypass restrictions. The research notes that even when platforms have ad policies against disinformation, enforcement is inconsistent.[54] Existing coordinated mechanisms for specific types of risks such as the Global Internet Forum to Counter Terrorism's (GIFCT), or mechanisms to connect regulators internationally such as the Global Online Safety Regulators Network (GOSRN) and the European Board for Digital Services, provide potential structures for information-sharing and coordination. However, they are either issue-specific, voluntary or advisory. They do not constitute a formalised, real-time coordination mechanism that links platforms, regulators and civil society specifically to address fast-moving conflict contexts.

Regulatory and legal context

Core international legal frameworks

The connection between international legal and normative frameworks and platforms underscores the need for effective measures to address conflict-related risks. International human rights law (IHRL) offers vital protections, such as the freedom of expression, freedom from discrimination, and the prohibition of incitement to violence or hatred. Article 19 in both the Universal Declaration of Human Rights (UNDHR) and the International Covenant on Civil and Political Rights (ICCPR) uphold the right to seek and share information.[55] The ICCPR specifies various legitimate restrictions (solely to safeguard the rights of others) to preserve national security, public order, public health or morals. Additionally, Article 20 of the ICCPR explicitly prohibits propaganda for war and advocacy of national, racial or religious hatred that incites discrimination or violence. While these obligations mainly concern states, private companies are urged to comply with similar standards.

Legal scholars have debated to what extent International Humanitarian Law (IHL) and International Criminal Law (ICL) may also apply to tech companies, regardless of geographic location. No social media platform provider has yet been held accountable in an international tribunal, as jurisdictional questions over corporate entities persist. However, there may be a legal basis for individuals within a company to be held accountable.[56] At the same time, the importance of social media companies in tracking and recording violations has grown substantially: evidence gathered from these platforms is frequently used to ensure accountability for serious offences such as war crimes and crimes against humanity. As a result, social media companies are under increasing pressure to manage harmful content to prevent any implication in international crimes.

Mechanisms and tools supporting accountability

The Berkeley Protocol is a global set of standards and methods for identifying, collecting, preserving, verifying, analysing and reporting digital open-source information when investigating alleged breaches of ICL, IHL and IHRL.[57] The Protocol launched on 1 December 2020 and was developed by the Human Rights Center at the University of California, Berkeley, and the United Nations Office of the High Commissioner for Human Rights (OHCHR). The framework is aimed at those investigating potential international human rights and criminal law infringements, guiding them in implementing best practices to generate reliable evidence. The Protocol is not a legal mechanism. However, it supports professionalising open-source investigations and ensuring that digital information is managed systematically, ethically and in a legally compliant manner. It can help produce reliable evidence to facilitate accountability and withstand international scrutiny.

UN-mandated bodies have also examined the role of online platforms in conflict. For example, the Independent International Fact-Finding Mission on Myanmar in 2018 made strong findings that Facebook significantly contributed to hateful incitement and violence. These (non-binding) findings are widely cited regarding the role of social media in the Rohingya crisis.[58] As a result, the Gambia has invoked Facebook in evidence-gathering efforts, using US court orders to obtain internal Facebook data to support the genocide case against Myanmar at the International Court of Justice (ICJ).[59] The US courts' 2021 ruling, largely based on the Stored Communications Act (SCA), should have supplied the critical evidence the Gambia sought. However, it is not publicly known to what extent Facebook passed on the requested information. The final judgment by the ICJ is pending.[60]

Corporate responsibilities under international normative frameworks

The UN Guiding Principles on Business and Human Rights (UNGP) serve as a global standard-setting framework to prevent infringements on human rights. They were endorsed in June 2011 by the UN Human Rights Council and published in 2012 by the OHCHR. The UNGP guide businesses, including tech companies, in mitigating human rights risks, including conflict-related risks, while conducting their operations.[61] Despite the UNGP urging businesses to conduct comprehensive assessments and stakeholder engagement, the non-binding nature leaves accountability dependent on state uptake and company goodwill.

Global and multi-stakeholder initiatives on digital content governance

Other global initiatives to address the discussed conflict-related risks include the UN Global Digital Compact, established on 22 September 2024. This established common standards to combat harmful digital content, advocating for adherence to privacy and freedom of expression while emphasising transparent moderation and collaboration with civil society.[62]

The Santa Clara Principles were established in May 2018 by a coalition of organisations, advocacy groups, and scholars in response to concerns about transparency and accountability in content moderation. The principles advocate for clear user notifications and effective appeals from platforms.[63]

Meanwhile, the Christchurch Call urges governments and technology companies to remove terrorist content, promoting prevention, rapid responses and cooperative legislative frameworks which uphold human rights.[64]

In 2022, Access Now (with input from multiple partner organisations such as ARTICLE 19 and the Center for Democracy and Technology) published a declaration of principles for content and platform governance in times of crisis.[65] This declaration outlines guidelines that assist platforms in safeguarding human rights across all phases of a crisis, aligning their actions with international human rights obligations.

National and regional legislation

Addressing conflict-related online risks requires more than voluntary platform action. It also needs binding regulatory measures that can protect users from conflict-related risks that may affect users outside of the conflict zone itself. Robust national and regional frameworks (such as the EU's DSA and the UK's OSA) can play a critical role in stemming the cross-border spread of disinformation, hate speech and other online risks originating from conflict zones. They can ensure some form of accountability for tech platforms that may otherwise operate with impunity. These laws can compel platforms to assess and mitigate systemic risks, respond rapidly to crises and provide transparency and data access for independent oversight. By doing so, they help safeguard vulnerable communities and uphold information integrity, although their ultimate impact depends on effective implementation and enforcement.

Regulating conflict-related online risks under the EU Digital Services Act

The EU's DSA establishes a regulatory framework for online intermediaries, placing heightened obligations on Very Large Online Platforms and Search Engines (VLOPSEs) especially in regard to systemic risks. Many conflict-related online risks outlined in this brief, including as illegal content (e.g. terrorist content, hate speech, incitement to violence), disinformation and foreign interference, and their algorithmic amplification, could qualify as "systemic risks" if they occur in a consistent or predictable way. This includes when they originate in conflict zones but affecting users within the EU.

The DSA obliges VLOPSEs to identify and assess (Article 34) and mitigate (Article 35) systemic risks. The European Commission, as the primary enforcer for VLOPSEs, can also mandate measures in ordinary circumstances under Articles 34-35, and in extraordinary circumstances under the crisis response mechanism (Article 36). By contrast, crisis protocols under Article 48 are explicitly voluntary. The European Commission may facilitate or encourage the draw-up, testing or adaptation of crisis protocols, but it cannot mandate them. Further, the DSA establishes new transparency and reporting obligations on all online platforms. These include annual transparency reports on content moderation, though VLOPSEs must meet more extensive requirements. These services must submit risk assessment reports (Article 34) and risk mitigation reports (Articles 35) on an annual basis to the Commission and national Digital Services Coordinators (DSCs).

Crisis Response Mechanism (Article 36): The Crisis Response Mechanism allows the Commission, on the recommendation of the European Board for Digital Services, to require VLOPSEs to assess their role in an ongoing crisis, and to take proportionate measures to prevent or limit such contributions. "Crisis" is defined as extraordinary circumstances posing a serious threat to public security or health. Recital 91 cites the examples of armed conflict, terrorism, natural disasters, pandemics and emerging conflicts. Measures may include (but are not limited to) adapting recommender systems, prioritising verified information or intensifying cooperation with trusted flaggers.

In practice, when a crisis is declared, the Board adopts a decision to act. This enables the European Commission to assess the functioning of services, order measures to prevent or limit their contribution to the threat, and obtain detailed information on the content in question, the implementation process, and the impact of the measures taken.[66] Any measures must be "strictly necessary, justified and proportionate" and may not exceed three months (Article 36(3) DSA). The Commission must publish crisis-response decisions, inform platforms immediately, grant the Board access to relevant information, and report annually to the European Parliament and the Council (Article 36(4), (11) DSA).

Crisis Protocols (Article 48) also enable the European Commission to initiate voluntary crisis protocols for "crisis situations… limited to extraordinary circumstances affecting public security or public health". These non-binding protocols allow the Commission to coordinate with VLOPSEs for actions such as designating crisis points of contact and

reallocating compliance resources. Advocates of crisis protocols and response mechanisms contend that these measures serve as an essential safeguard for fundamental rights during crises involving public health or security threats: for instance, by prominently displaying verified information on a service's front page to curb misinformation.[67]

Some critics including members of civil society[68] argue that the lack of clarity around the implementation of these emergency measures threatens the rule of law. This is particularly true when it is unclear what constitutes a crisis event and who decides when an emergency is declared. Given this, ISD and others have called on the Commission to clarify whether it will interpret a crisis event as defined in the DSA as a "state of emergency" in international human rights law.[69]

Data Access and Scrutiny (Article 40): Article 40 enables researchers to access two types of platform data through different means: a) public data (via APIs or data libraries) and b) non-public data, accessible to researchers vetted by national DSCs. These provisions are a critical tool for researching and addressing conflict-related online risks. Access to granular data on content removal, algorithmic ranking and cross-border content flows would enable such actors to serve as an early warning system for human rights violations, including those affecting vulnerable communities within the EU. Together with independent audit requirements (Article 37), this also supports accountability by verifying whether platforms enforce their own policies effectively, not merely whether such policies exist. When combined with systemic risk obligations under Articles 34–35, these provisions could help EU regulators detect and mitigate conflict-linked risks in real time, protect diaspora communities, and strengthen information integrity during crises.

Safeguards Against Over-Moderation Under the DSA: The DSA embeds robust safeguards to prevent over-moderation and uphold freedom of expression in conflict-related contexts. Articles 14, 17, 20 and 21 require platforms to apply their terms impartially, provide public reasoning for content removals, and ensure users have access to both internal complaint systems and out-of-court dispute resolution. Platforms must also incorporate assessments of potential risks to fundamental rights (including the risk of unjustified removals) within their systemic risk management processes (Article 34) and provide transparent reporting of moderation actions (Articles 14 and 19). These protections are critical during times of conflict when rapid content removal risks suppressing legitimate documentation, human rights reporting or political speech from vulnerable communities. However, while these protections are strong on paper, their effectiveness depends on consistent and timely enforcement. This again underscores the need for active regulatory oversight and independent scrutiny.

The UK Online Safety Act and its relevance to transnational conflict-related harms

The UK's OSA establishes a comprehensive regulatory framework to address a wide range of online harms, placing statutory duties on providers of regulated user-to-user and search services. As designated regulator, Ofcom is responsible for issuing and enforcing Codes of Practice that ensure compliance with these duties, including freedom of expression and privacy in line with the UK's human rights obligations.

While the OSA's primary aim is to protect UK residents, its provisions are highly relevant for harms that originate in conflict zones abroad but manifest in the UK's information environment. These include such as foreign interference, illegal content posing risks to minority or diaspora populations, or inciting violence. The Act applies to content accessible to UK users regardless of its geographic origin, making it a potential tool to mitigate cross-border harms stemming from armed conflicts.

Conflict-relevant priority illegal harms addressed under the OSA include:

- **Foreign interference** – The OSA lists Foreign Interference offences in Schedule 7 (priority offences for terrorism and national security, as defined in Section 13 of the National Security Act 2023). Ofcom is supported by an Online Information Advisory Committee (Section 152) and has media literacy duties to reduce public exposure to manipulation and propaganda.[85] Disinformation is only included in this when attributed to state-linked or foreign interference.

- **Hate offences** – Offences under the Public Order Act 1986 and the Crime and Disorder Act 1998 are designated as "priority offences" in Schedule 7. This refers specifically to "stirring up of racial hatred; stirring up of hatred on the basis of religion or sexual orientation".

- **Terrorism offences**: Schedule 5 of the OSA explicitly lists terrorism offences, which include offences relating to 'proscribed organisations', information likely to be of use to a terrorist, training for terrorism, encouraging terrorism or disseminating terrorist materials, and financing of terrorism.[86]

DSA Case study
Risk assessment reports submitted by VLOPSEs

As part of this report, ISD analysed the public risk assessment reports published in 2023 and 2024 by VLOPSEs under the DSA. While all the risk assessment reports from this period were reviewed, only selected VLOPSEs' reports are presented here based on their inclusion of relevant elements for conflict-related risks. The sample covers reports from: AliExpress (2024), Amazon (2023, 2024), Apple (2023, 2024), Bing (2023, 2024), Booking.com (2023, 2024), Facebook (2024), Google (2023, 2024), Instagram (2024), LinkedIn (2023, 2024), Pinterest (2023, 2024), Snapchat (2023), TikTok (2023, 2024), Wikipedia (2023), X (2023) and Zalando (2023, 2024).

The aim was to examine how and to what extent these platforms interpret and operationalise their obligations under Articles 34–35 on identifying, assessing and mitigating systemic risks in relation to armed conflict and crisis situations. Using a structured keyword search for references to "(armed) conflict," "war(s)," or "crisis/crises", ISD analysed how platforms frame conflict-linked risks and the measures they report to address them.

Key cross-platform findings: Several services have operationalised crisis protocols as standing playbooks, while others provide more generic references to crisis preparedness. This difference may reflect operational maturity or the degree to which conflict-related risks are relevant to a specific type of service.

Varied approaches to conflict in platforms' risk assessments: Booking.com removed "war crimes/genocide" from its risk matrix in 2024 on the grounds that it was "insufficiently material" to its operations[70]; LinkedIn stated that its professional orientation and slower virality meant crises were not likely to pose public-security risks.[71] By contrast, TikTok categorised armed conflict as a priority systemic risk and provided the most detailed crisis-response disclosures.[72] Apple's 2023 report noted that App Store Guidelines prohibit apps seeking to profit from violent conflicts or terrorist attacks,[73] while its 2024 risk assessment no longer specifically mention conflict-linked risks.[74]

TikTok provides the most detailed and operationally specific crisis-response measures. Most other platforms disclose only vague or high-level protocols, revealing a significant transparency and consistency gap in how crises are covered in platforms DSA risk assessments. TikTok describes a six-phase crisis management system, provides activation timelines down to the minute following Hamas' attack on Israel on October 7, and references the launch of a 200-person command centre with content moderation in Arabic and Hebrew.[75] Shortly after the beginning of the war, TikTok suspended recommendation eligibility for short-video and LIVE content originating from Israel or Palestine in the EU. This represents the only origin-based suspension disclosed in the analysed platforms' risk assessment reports in response to a conflict situation.[76] While other platforms have implemented visibility-limiting mitigations, this is the only concrete reference to a reach-limiting measure with an explicit geographical scope. TikTok references the need to balance risk mitigations with protections for freedom of expression elsewhere in its risk assessment. However, it does not explicitly describe how freedom of expression is considered in the context of such a blanket approach to reducing the visibility of content on a geographical basis.

Other platforms provide less granularity or operational detail, for example:

- **Instagram** and **Facebook** describe crisis response protocols triggered by external events (e.g. wars, protests and elections, threat-signal sharing with governments, and periodic country prioritisation).[77] However, their risk assessment reports do not refer to critical crisis protocol elements. These include activation thresholds, the rules or settings applied (manually or automatically) by services to their recommendation systems during a crisis, or the conditions to end crisis-activation measures in their risk assessment reports.

- **Google** sets out a "sensitive events" policy in Ads (prohibiting exploitative ads during conflicts). The platform notes that Google Play can temporarily pause new user reviews during sensitive events. However, it does not provide details on recommender systems or (de)activation thresholds for YouTube.[78]

- **Bing** refers to "sensitive queries" management and crisis-related signals to boost authoritative sources in search results without specifying thresholds, coverage levels or evaluation methods.[79] The search engine noted participation in EU crisis working groups during the COVID-19 pandemic and full-scale invasion of Ukraine in the 2023 report. In 2024, it committed to "further enhance crisis and rapid response protocols," treating elections, armed conflicts, and AI misuse as key global risks.[80]

- **AliExpress** framed risks broadly, prohibiting content that endangers sovereignty, advocates wars or disseminates violence. However, it did not offer evidence of conflict-specific activation mechanisms.[81]

Other platforms made no references to crisis response in their risk assessment reports.

In the context of crisis situations, most platforms do not specifically mention language capabilities of crisis response or content moderation teams. Facebook, Instagram and Google refer to "multilingual moderation" but do not specify coverage by language.[82] Most risk assessment reports do not specifically mention whether moderation teams can scale or adapt reactively to conflict-relevant languages. They also do not cross-reference DSA transparency reports, which disclose the number of moderators per official EU language. As mentioned above, TikTok describes its capabilities for "conflict-relevant" languages.

Some platforms report metrics such as the volume of content removals, user engagement with safety tools such as fact-checking labels, information hubs and reporting functions. These figures illustrate the scale of enforcement or feature uptake. However, they do not provide outcome-based indicators. This poses challenge for regulators when assessing the actual effectiveness, or the proportionality of crisis responses in mitigating or reducing risks. TikTok reports the removal of 3.1 million videos and 140,000 livestreams linked to the Israel–Hamas war. X highlights the scale of Community Notes as an output metric. None of the services or platforms provide systematic reporting on impact indicators, such as whether the removal of illegal or Terms of Service (ToS)–violating content reduced engagement or limited its reach. This lack of evidence makes it difficult for regulators to assess the proportionality or effectiveness of mitigation measures.

Regulatory enforcement: In December 2023, the European Commission opened its first formal proceeding against X, based on suspected infringements of Article 34 and Article 35 in terms of systemic risk management, illegal content moderation, dark patterns, advertising transparency and data access for researchers (Art. 40). This followed widespread disinformation during the Israel–Hamas war.[83] The European Commission also issued conflict-triggered requests for information (RFIs) about Meta and TikTok's crisis response efforts; these were explicitly linked to the spread of illegal content and disinformation at the start of the Israel-Hamas war. Those RFIs were later followed by formal DSA proceedings in 2024 on broader systemic-risk and transparency grounds.[84]

Analysis of the reports suggests that platforms have partially developed formalised crisis protocols but lack transparency on key aspects of their operations. Platforms disclose little about how recommender systems are adapted under crisis activation, how long interventions last, or what safeguards are applied to prevent the over-removal of legal and/or ToS-compliant content. When it comes to assessing the effectiveness of employed measures, metrics of effectiveness, if mentioned, are skewed toward outputs (e.g. the number of removals) rather than outcomes (e.g. metrics demonstrating a reduction in harmful reach or time-to-mitigation).

The European Commission's RFIs and recent proceedings demonstrate that conflict-linked crises are already a focus of DSA enforcement. To ensure that risk assessments fulfil their intended purpose, future reporting should move beyond process descriptions and output metrics towards outcome-based metrics, language-specific disclosures and auditable records of adjustments to recommendation systems. Without these, there is a risk that reporting remains primarily procedural rather than providing meaningful accountability for systemic risks during times of conflict.

In the OSA, Section 59 establishes providers' duties in relation to "priority illegal content". Schedule 7 specifies the criminal offences that fall within this category (see above), defining its scope and ensuring clarity on which offences trigger heightened duties. Hence, platforms are placed under clear and proactive obligations to reduce the risks of priority illegal offences occurring.

Similarly to the EU's DSA, the OSA requires services to conduct regular risk assessments to understand how design features and functionalities may facilitate the dissemination of such material. They must then implement proactive systems and processes to reduce the likelihood of users encountering priority illegal content (for example, through moderation tools, effective reporting and escalation procedures, and adjustments to recommender systems). Where such content is identified, it must be removed swiftly. Platforms are expected to take proportionate steps to mitigate broader risks, such as improving account verification or investing in language-specific moderation. These duties are reinforced by transparency requirements: platforms are required to report publicly on how they address priority illegal content and by Ofcom's powers to audit, compel information and enforce compliance.

Services are designated by Ofcom into categories according to their size and risk profile. Category 1 services, the largest user-to-user platforms, are subject to the most extensive obligations. These include enhanced duties to protect democratic and journalistic content, produce regular transparency reports, and provide user empowerment tools. They also have duties for illegal priority content and harms to children (minors). Category 2 services carry more limited duties: Category 2A covers major search engines while Category 2B applies to medium-sized user-to-user platforms. These services must still assess and mitigate risks of illegal content and harms to children. However, they are not bound by the additional obligations reserved for Category 1 platforms. Failure to meet these obligations can result in significant sanctions, including fines of up to £18 million or 10% of global turnover. In serious cases, sanctions may include liability for senior managers.

The OSA recognises the role of platform design and business models in amplifying illegal content and content harmful to children. It requires providers to assess risks from functionalities such as recommender systems and engagement-driven design (Sections 9, 12, 27, 29, and 208), as well as cumulative harms from repeated exposure, as set out in Ofcom's Risk Assessment Guidance. Transparency requirements in Chapter 6 (especially Section 65 and Schedule 8) require Category 1, Category 2A and 2B services to report on automated systems, including recommender systems and monetisation practices. These reporting obligations sit alongside Ofcom's Illegal Content Codes of Practice (User-to-User, Section E) and Children's Safety Codes; these also provide guidance on how services should address risks linked to design features including recommender systems and engagement-driven models.

The OSA embeds some safeguards against over-moderation. Providers must protect freedom of expression and privacy when implementing safety measures (Sections 22 and 33). Category 1 services must also protect content of democratic importance (Section 18) and journalistic content (Section 19). Providers must maintain complaints procedures to address wrongful removal (Sections 21 and 32) and ensure any use of proactive detection technologies is proportionate and rights-respecting; this is reflected in Ofcom's Illegal Content Codes of Practice (specifically, the recommendations under 'proactive technology') and Children's Safety Codes.

For conflict-related harms with transnational dimensions, the OSA offers an important framework. Ofcom could further strengthen its approach by integrating scenario-based risk monitoring for conflicts. Across recent incidents (especially anti-migrant violence such as the 2024 Southport riots) the lack of a formal crisis protocol in the OSA meant that platforms were neither required nor prepared to respond swiftly to rapidly evolving threats.[87] Crisis response mechanisms, as seen under the DSA, are notably absent from the OSA. They should be urgently integrated to close this critical gap.

Conclusion

As the number and complexity of armed conflicts rise, online spaces have become both a mirror of and a catalyst for conflict dynamics.[88] Risks originating in conflict zones do not stop at national borders. They reverberate across diasporas, polarise democratic societies, and undermine trust in institutions far from the physical sites of violence.

Efforts to mitigate conflict-related online risks have grown across the international humanitarian, multilateral and digital rights fields as demonstrated by initiatives such as the Santa Clara Principles, Berkeley Protocol, the Global Digital Compact and the UNGP which establish global standards. These efforts span the monitoring of online risks as well as an accountability framework for businesses and platforms themselves. The UNGPs marked a particularly significant step in clarifying corporate responsibilities to prevent and address threats to human rights including in fragile and conflict-affected contexts. However, because they remain voluntary, their effectiveness depends heavily on political and corporate will, adherence to norms, and the willingness of private companies to act.

There are lessons to be drawn from efforts to combat TVEC, where more structured cooperation mechanisms have been in place for several years. Initiatives such as the Christchurch Call and GIFCT illustrate how voluntary cooperation can be mobilised in moments of crisis. These examples highlight the potential of cross-platform coordination, information sharing and the need to ensure civil society participation.

In parallel, regulatory frameworks at the national and regional level have established a range of legal obligations on online platforms and service providers. The EU's DSA and the UK's OSA were not designed with conflict-related risks as their primary focus. Nonetheless, they contain important tools for mitigating these risks. The EU DSA establishes obligations including systemic risk assessments (Articles 34–35), a crisis response mechanism (Article 36), and enforceable transparency and researcher access requirements (Articles 40–42). These provisions allow regulators to mandate proportionate measures in extraordinary circumstances including armed conflict. The UK OSA, while not containing a dedicated crisis mechanism, imposes statutory duties to manage priority illegal content and requires platforms to conduct risk assessments and comply with Ofcom's Codes of Practice which cover several conflict-related harms.

These measures can help stem the spread of a wide range of online risks, from hate speech and the incitement to violence, TVEC, disinformation, FIMI, and GBV, as well as mitigate risks from over-moderation and content suppression. By encouraging a more structured and considered approach to managing online risks, such approaches can also help to encourage more responsible practices and design choices by global companies that play crucial roles in the information environment.

Despite these advances, significant gaps persist:

- **Reliance on voluntary cooperation**: Outside national and regional regulated jurisdictions, international efforts remain fragmented and largely dependent on goodwill from the private sector.
- **Evidence preservation**: The over-removal of content continues to undermine efforts to preserve and document evidence of human rights abuses. There are only a small number of typically independent mechanisms in place to ensure preservation.
- **Financial incentives**: Platforms' advertising and engagement-based business models can create financial rewards for divisive, sensational or misleading conflict-related content that generates high levels of engagement.
- **Stakeholder consultation**: Local actors, humanitarian organisations and diaspora communities are too rarely consulted as experts and affected communities to highlight emerging risks, provide vital contextual, and cultural or language expertise.
- **Normative gaps**: While the UNGPs provide an important global framework clarifying corporate responsibilities, they are non-binding and inconsistently applied even by companies that are signatories. This limits their impact in conflict settings.

Addressing these gaps will require coordinated action between governments, regulators, international organisations, civil society and the private sector. In doing so, democratic governments can strengthen both domestic information integrity and global resilience to conflict-related online risks.

Endnotes

1 Rustad, S. (2025). Conflict Trends: A Global Overview, 1946–2024. PRIO Paper. Retrieved from: https://www.prio.org/publications/14453.

2 In this brief we use the term "illegal hate speech" to align with the EU Digital Services Act (DSA), where it constitutes a category of illegal content. The DSA itself does not create a new definition, but refers to existing EU and Member State laws, most notably the 2008 Framework Decision on combating racism and xenophobia. This should be distinguished from other jurisdictions: in the UK, hate offences fall under the category of "priority illegal content" in the Online Safety Act; in Australia, relevant provisions are spread across online safety and anti-discrimination legislation; while in other jurisdictions, hate speech can be protected by constitutions and only be unlawful where it falls into narrower categories such as incitement or true threats.

3 Agarwal, A. & Latonero, M. (2021). Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar. Retrieved from: https://www.hks.harvard.edu/centers/carr/publications/human-rights-impact-assessments-ai-learning-facebooks-failure-myanmar and De Guzman, C. (28 September 2022). Meta's Facebook Algorithms 'Proactively' Promoted Violence Against the Rohingya, New Amnesty International Report Asserts. Retrieved from: https://time.com/6217730/myanmar-meta-rohingya-facebook/.

4 ISD. Effectiveness of the Sanctions on Russian State-Affiliated Media in the EU:
An Investigation into Website Traffic & Possible Circumvention Methods (6 October 2022).
Available at: https://www.isdglobal.org/wp-content/uploads/2022/10/Effectiveness-of-the-sanctions-on-Russian-state-media-1.pdf

5 Allen, C. (19 April 2022). Facebook's Content Moderation Failures in Ethiopia. Retrieved from: https://www.cfr.org/blog/facebooks-content-moderation-failures-ethiopia and Nicholas, G. & Bhatia, A. (2023, May 23). The Dire Defect of 'Multilingual' AI Content Moderation: https://www.wired.com/story/content-moderation-language-artificial-intelligence/

6 HCR. A practical toolkit to strengthen information integrity on digital platforms.
Retrieved from: https://www.unhcr.org/handbooks/informationintegrity/.

7 Collective service. For a community-led response. Retrieved from: https://www.rcce-collective.net/.

8 EDMO. EDMO Hubs. Retrieved from: https://edmo.eu/about-us/edmo-hubs/.

9 EEAS. (2019). Factsheet: Rapid Alert System. Retrieved from: https://www.eeas.europa.eu/node/59644_en.

10 Centre for humdata. (21 October 2021). The OCHA Data Responsibility Guidelines.
Retrieved from: https://centre.humdata.org/the-ocha-data-responsibility-guidelines/.

11 Rustad, S. (2025). Conflict Trends: A Global Overview, 1946–2024. PRIO Paper. Retrieved from: https://www.prio.org/publications/14453

12 ICRC. (2025). Addressing harmful information in conflict settings: A Response Framework for Humanitarian Organizations. Retrieved from: shop.icrc.org/addressing-harmful-information-in-conflict-settings-a-response-framework-for-humanitarian-organizations-pdf-en.html.

13 UN. (2019). The UN Strategy and Plan of Action.
Retrieved from: https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech.

14 UN Security Council. Resolution 1624 (2005). Retrieved from: https://digitallibrary.un.org/record/556538?v=pdf.

15 UN Security Council. Resolution 2354 (2017). Retrieved from: https://digitallibrary.un.org/record/1289209?v=pdf.

16 EU. Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online. Retrieved from: https://eur-lex.europa.eu/eli/reg/2021/784/oj/eng.

17 UN General Assembly. Resolution adopted by the General Assembly on 24 December 2021. 76/227 Countering disinformation for the promotion and protection of human rights and fundamental freedoms. Retrieved from: https://www.un.org/en/ga/76/resolutions.shtml.

18 EEAS. (2025). Information Integrity and Countering Foreign Information Manipulation & Interference.
Retrieved from: Information Integrity and Countering Foreign Information Manipulation & Interference (FIMI) | EEAS.

19 UN. General recommendation No. 35 (2017) on gender-based violence against women, updating general recommendation No. 19 (1992). Retrieved from: https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-recommendation-no-35-2017-gender-based.

20 ICRC. (2024). Online violence: real life impacts on women and girls in humanitarian settings. Retrieved from: https://blogs.icrc.org/law-and-policy/2024/01/04/online-violence-real-life-impacts-women-girls-humanitarian-settings/.

21 Guess, A. et al. (2023). Reshares on social media amplify political news but do not detectably affect beliefs or opinions.
Retrieved from: https://www.science.org/doi/10.1126/science.add8424.

22 Schirch, L. (2021). Social Media Impacts on Conflict and Democracy. The Tectonic Shift.
Retrieved from: https://www.routledge.com/Social-Media-Impacts-on-Conflict-and-Democracy-The-Techtonic-Shift/Schirch/p/book/9780367541057?srsltid=AfmBOorTQljyOg7Yj-eZnieeKRDYmoscd5tZvTYb-rFCAk15aY2c5G85.

23 UNHRC. (2018). Report of the independent international fact-finding mission on Myanmar. Retrieved from: https://www.ohchr.org/sites/default/files/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf. And: Amnesty International. (2022). Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations. Retrieved from: https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/.

24 Crystal, C. (2018). Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech. Retrieved from: https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en. Demony, C. (19 February 2025). Israel-Gaza war fuels record level of anti-Muslim hatred in Britain, monitoring groups says. Retrieved from: https://www.reuters.com/world/uk/israel-gaza-war-fuels-record-level-anti-muslim-hatred-britain-monitoring-group-2025-02-19/. And: CST. (2025). Antisemitic Incidents Report 2024. Retrieved from: https://cst.org.uk/news/blog/2025/02/12/antisemitic-incidents-report-2024.

25 ISD. (2025). Conflict Amplified: Disinformation and Hate in the Israel-Hamas War. Retrieved from: Conflict-amplified_Disinformation-and-Hate-in-the-Israel-Hamas-War.pdf.

26 Meixler, E. (13 March 2018). UN Fact Finders Say Facebook Played a 'Determing' Role in Violence Against the Rohingya. Retrieved from: https://time.com/5197039/un-facebook-myanmar-rohingya-violence/.

27 Business & Human Rights Resource Centre. (27 May 2025). Ethiopia: Social media platforms "Failed to adequately moderate genocidal content" during Tigray war. Retrieved from: https://www.business-humanrights.org/en/latest-news/ethiopia-social-media-platforms-failed-to-adequately-moderate-genocidal-content-during-tigray-war-study-finds-x-formerly-twitter-did-not-respond/.

28 Pírková, E. & Fatafta, M. (2022). Content governance in times of crisis: How platforms can protect human rights. Retrieved from: https://www.accessnow.org/publication/new-content-governance-in-crises-declaration/.

29 Alimardani, M., Gregory, S. (2025). Iran-Israel AI War Propaganda Is a Warning to the World. Carnegie Endowment for International Peace. Retrieved from: Iran-Israel AI War Propaganda Is a Warning to the World | Carnegie Endowment for International Peace.

30 Baig, R. (2025). Fact check: Viral Evin prison blast video is likely AI fake. Retrieved from: Fact check: Viral Evin prison blast video is likely AI fake – DW – 06/27/2025.

31 European Digital Media Observatory. (2025). The First AI War: How the Iran-Israel Conflict Became a Battlefield for Generative Misinformation. Retrieved from: The First AI War: How the Iran-Israel Conflict Became a Battlefield for Generative Misinformation – EDMO.

32 ISD. (2025). Missiles and misinformation: false claims about the India-Pakistan clashes reach millions on X. Retrieved from: Missiles and misinformation: false claims about the India-Pakistan clashes reach millions on X - ISD.

33 International Federation of Red Cross and Red Crescent Societies and the International Committee of the Red Cross. (2024). Protecting civilians and other protected persons and objects against cyber and information operations during armed conflict. Background document. Retrieved from: 34IC-Background-doc-Cyber-EN.pdf.

34 ISD. (2025). Conflict Amplified: Disinformation and Hate in the Israel-Hamas War. Retrieved from: Conflict-amplified_Disinformation-and-Hate-in-the-Israel-Hamas-War.pdf.

35 Furstenberg, S., Michaelsen, M., Anstis, S. (2025). Transnational repression of human rights defenders: The impacts on civic space and the responsibility of host states. Retrieved from: Transnational repression of human rights defenders: The impacts on civic space and the responsibility of host states | Think Tank | European Parliament.

36 BBC. (24 May 2024). Why Eritreans are at war with each other around the world. Retrieved from: Eritrean Independence Day: Why the diaspora is at war with itself.

37 ISD. (2025). Conflict Amplified: Disinformation and Hate in the Israel-Hamas War. Retrieved from: https://www.isdglobal.org/isd-publications/conflict-amplified-disinformation-and-hate-in-the-israel-hamas-war/.

38 Ibid.

39 United Nations. (1966). International Covenant on Civil and Political Rights (ICCPR). Retrieved from: ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

40 Fundación Multitudes. (2024). Online Gender-Based Violence against Women* Environmental and Human Rights Defenders in Latin America. The cases of Ecuador, Peru and Colombia. Published by Heinrich Böll Stiftung. Retrieved from: Online Gender-Based Violence against Women* Environmental and Human Rights Defenders in Latin America | Heinrich Böll Stiftung.

41 EEAS. (2022). US/EU: Joint Statement on Protecting Human Rights Defenders Online. Retrieved from: US/EU: Joint Statement on Protecting Human Rights Defenders Online | EEAS.

42 Article 19. (2023). Content Moderation and freedom of expression handbook. Retrieved from: SM4P-Content-moderation-handbook-9-Aug-final.pdf.

43 Mnemonic. Caught in the Net: The Impact of Extremist Speech Regulations on Human Rights Content. Retrieved from: mnemonic.org/en/content-moderation/caught-in-the-net-the-impact-of-extremist-speech-regulations-on-human-rights-content/.

44 Mingo, M. (2024). Why Meta needs to implement these Oversight Board recommendations. Retrieved from: mnemonic.org/en/content-moderation/why-meta-needs-to-implement-these-oversight-board-recommendations/.

45 Ridsdale, C. (2022). Fueling the Flames: Online Social Platform Responsibility for Harmful Content in Conflict Zones. Retrieved from: https://www.mcgill.ca/humanrights/files/humanrights/charlotte_ridsdale_-_fueling_the_flames_-_online_social_platform_responsibility_for_harmful_content_in_conflict_zones.pdf.

46 Human Rights Watch. (2023). Meta's Broken Promises. Systemic Censorship of Palestine Content on Instagram and Facebook. Retrieved from: https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and.

47 Fatafta, M. (19 February 2024). It's not a glitch: how Meta systematically censors Palestinian voices. Retrieved from: https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/.

48 Business & Human Rights Resource Centre. (15 October 2023). Meta faces users allegations of censorship Pro-Palestinian voices on Instagram. Retrieved from: https://www.business-humanrights.org/en/latest-news/user-meta-faces-users-allegations-of-censorship-pro-palestinian-voices-on-instagram/.

49 Syrian Archive. (n.d.) How we do it. Retrieved from: https://syrianarchive.org/
50 ISD. (2022). 'Suggested for You': Understanding How Algorithmic Ranking Practices Affect Online Discourses and Assessing Proposed Alternatives. Retrieved from: https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/.
51 Reset Tech. (2024). Verified Disinformation: How X Profits from the Rise of a Pro-Kremlin Network. https://www.reset.tech/resources/verified-disinformation-research-report-reset-tech-2024_web.pdf.
52 Ibid.
53 Ibid.
54 Ibid.
55 United Nations. (n.d.). Universal Declaration of Human Rights. Retrieved from: https://www.un.org/en/about-us/universal-declaration-of-human-rights. And: Office of the United Nations High Commissioner for Human Rights (OHCHR). (n.d.). International Covenant on Civil and Political Rights. United Nations. Retrieved from: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.
56 Dardari, A., Levsen, N., Setian, A., & Peake, J. (2022). Social media content moderation and international human rights law. The example of the Nagorno-Karabakh/Artsakh Conflict. Promise Institute for Human Rights, UCLA School of Law. Retrieved from: https://promiseinstitute.law.ucla.edu/wp-content/uploads/2022/05/Social-Media-Content-Moderation-and-Internationals-Human-RIghts-Law.pdf
57 Office of the United Nations High Commissioner for Human Rights (OHCHR). (n.d.). Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law. United Nations. Retrieved from: https://www.ohchr.org/en/publications/policy-and-methodological-publications/berkeley-protocol-digital-open-source
58 Hunt, K. (13 March 2018). U.N. Fact Finders Say Facebook Played a 'Determining' Role in Violence Against the Rohingya. Time. Retrieved from: https://time.com/5197039/un-facebook-myanmar-rohingya-violence/
59 Keenan, E. (23 September 2021). Q&A on Court Ordering Facebook to Disclose Content on Myanmar Genocide. Just Security. Retrieved from: https://www.justsecurity.org/78358/qa-on-court-ordering-facebook-to-disclose-content-on-myanmar-genocide
60 International Court of Justice (2025). Pending cases. Retrieved from: https://icj-cij.org/pending-cases
61 Office of the United Nations High Commissioner for Human Rights (OHCHR). (2011). Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. United Nations. Retrieved from: https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights.
62 United Nations. (n.d.). Global Digital Compact. Retrieved from: https://www.un.org/digital-emerging-technologies/global-digital-compact.
63 Santa Clara Principles. (n.d.). The Santa Clara Principles on Transparency and Accountability in Content Moderation. Retrieved from: https://santaclaraprinciples.org/.
64 Christchurch Call. (n.d.). Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online. Retrieved from: https://www.christchurchcall.org/.
65 Access Now. (2022). Declaration of principles for content and platform governance in times of crisis. Retrieved from: https://www.accessnow.org/wp-content/uploads/2022/11/Declaration-of-principles-for-content-and-platform-governance-in-times-of-crisis.pdf.
66 Sekwenz, M. & Gsenger, R. (2025). The Digital Services Act: Online Risks, Transparency and Data Access. In: Digital Descade: How the EU Shapes Digitalisation Research. Retrieved from: https://doi.org/10.5771/9783748943990%E2%80%91115.
67 Schwieter, C. (2022). Online crisis protocols – Expanding the regulatory toolbox to safeguard democracy during crisis. Retrieved from: https://www.isdglobal.org/wp-content/uploads/2022/12/Online-Crisis-Protocols-%E2%80%93-Expanding-the-Regulatory-Toolbox-to-Safeguard-Democracy-During-Crises.pdf.
68 Article 19. (13 April 2022). EU: Digital Services Act crisis response mechanism must honour human rights. Retrieved from: https://www.article19.org/resources/eu-digital-services-act-crisis-response-must-respect-human-rights/.
69 Schwieter, C. (2022). Online crisis protocols – Expanding the regulatory toolbox to safeguard democracy during crisis. Retrieved from: https://www.isdglobal.org/wp-content/uploads/2022/12/Online-Crisis-Protocols-%E2%80%93-Expanding-the-Regulatory-Toolbox-to-Safeguard-Democracy-During-Crises.pdf.
70 Booking.com. (2024). Mitigation Measures Completion Report conducted by Booking.com B.V. under the Digital Services Act, Retrieved from: https://r-xx.bstatic.com/data/mobile/8c247ca8-378a-434b-95dc-d233ee23430c.pdf-. p. 16.
71 LinkedIn (2024). Systemic Risk Assessment. Retrieved from: An update on compliance with the Digital Services Act. pp. 40.
72 TikTok. (2023). DSA Risk Assessment Report. Retrieved from: https://panoptykon.org/sites/default/files/2025-01/tiktok-dsa-risk-assessment-report-2023.pdf. p. 27.
73 Apple App Store (2023). Report on Risk Assessment and Risk Mitigation Measures. Retrieved from: 20232808_app-store_risk-assessment-report_non-confidential.pdf. p. 44, 54.
74 Apple App Store (2024). Second Report on Risk Assessment and Risk Mitigation Measures. Retrieved from: 20241212_app-store_risk-assessment-report_non-confidential.pdf.
75 TikTok. (2024). DSA Risk Assessment Report 2024. Retrieved from: https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwlY_fjulyhwzuhy[/ljhwZthlaukjlkulzlp/DSA/TikTok_DSA_Risk_Assessment_Report_2024.pdf. p. 28.
76 Ibid.

77 Meta. (2024). Regulation (EU) 2022/2065 Digital Services Act (DSA). Systemic Risk Assessment and Mitigation Report for Instagram. Retrieved from: https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2024-8-28_2024-8-28_en_v1.pdf. p. 23.

78 Google. (2024). Report of Systemic Risk Assessments. Retrieved from: https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2024-8-28_2024-8-28_en_v1.pdf pp. 22., p. 123.

79 Bing. (2024). Retrieved from: August-2024-Microsoft-Bing-Systemic-Risk-Assessment-Report-EU-Digital-Services-Act.pdf.

80 Ibid.

81 AliExpress. (2024). Digital Services Act. Risk Assessment and Mitigation Measures Report 2024. Retrieved from: aliexpress.com/p/transparencycenter/mitigationReport.html.

82 Google. (2024). Report of Systemic Risk Assessments. Retrieved from: https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2024-8-28_2024-8-28_en_v1.pdf, and Meta. (2024). Regulation (EU) 2022/2065 Digital Services Act (DSA). Systemic Risk Assessment and Mitigation Report for Instagram. Retrieved from: https://storage.googleapis.com/transparencyreport/report-downloads/dsa-risk-assessment_2024-8-28_2024-8-28_en_v1.pdf.

83 Commission opens formal proceedings against X under the Digital Services Act (Press release, December 2023); Digitale Strategie Europa, "DSA Proceedings against X," 2023. Retrieved from: Commission opens formal proceedings against X under the Digital Services Act | Shaping Europe's digital future.

84 Commission opens formal proceedings against Meta under the Digital Services Act (Press release, April 2024); Society for Computers & Law, "DSA enforcement update: Meta under investigation,". Retrieved from: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664.

85 Online Information Advisory Committee (10 June 2025), Ofcom. Retrieved from: https://www.ofcom.org.uk/about-ofcom/structure-and-leadership/advisory-committee-on-disinformation-and-misinformation.

86 Ofcom (n.d.). Overview of Illegal Harms. Retrieved from: https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/overview-of-illegal-harms.pdf?v=390985.

87 ISD. (15 August 2024). After Southport: Policy responses to far-right extremism. Retrieved from: https://www.isdglobal.org/digital_dispatches/after-southport-policy-responses-to-farright-extremism/.

88 Rustad, S. (2025), Conflict Trends: A Global Overview, 1946–2024. PRIO Paper. Retrieved from: https://www.prio.org/publications/14453.

ISD | Institute for Strategic Dialogue

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org