unesco

UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

PROTECTING CRITICAL VOICES

Guidance for Human Rights Impact Assessment on Digital Platforms

Published in 2025 by the United Nations Educational, Scientific and Cultural Organization (UNESCO), 7, place de Fontenoy, 75007 Paris, France and the Office of the High Commissioner for Human Rights (OHCHR), Palais Wilson, Rue des Pâquis 52, 1201, Geneva, Switzerland.

ISBN 978-92-3-100837-5

DOI https://doi.org/10.58338/YOXE5767

This publication should be cited as follows: UNESCO, OHCHR. 2025. *Protecting Critical Voices: Guidance for Human Rights Impact Assessment on Digital Platforms*.

Editors: UNESCO and OHCHR

Contributing Authors: UNESCO and OHCHR

Cover, graphic design and typset: Luiza Maximo

Printed by: UNESCO

Printed in Paris

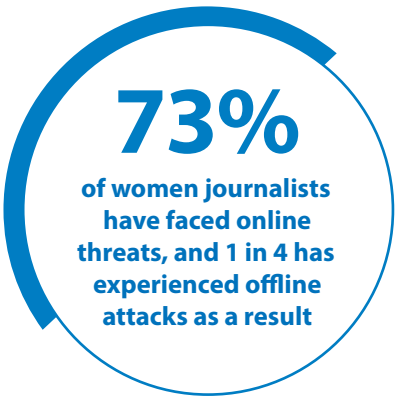
SHORT SUMMARY

Protecting critical voices in the digital age

As providers of information and communication channels for billions of people around the world, digital platforms profoundly impact the rights to freedom of expression and participation, by offering a space for public discourse and civic engagement. Online spaces and digital platforms have become a central arena for journalists and human rights defenders to perform their roles. At the same time, these are spaces where critical voices are exposed to threats and attacks that may result in censorship or self-censorship, impacting the individuals concerned, but also more broadly impacting societies' access to information and trust in institutions.

This Guidance aims to help companies in identifying, assessing and responding to risks to human rights associated with digital platforms, particularly in relation to the protection of journalists and human rights defenders.

UNESCO and the Office of the United Nations High Commissioner for Human Rights (OHCHR) developed this Guidance, building on their engagement with a range of stakeholders involved in the protection of critical voices and providing guidance to companies on measures relevant to safeguard human rights in the digital sphere. It rests on the recognition that collaboration between stakeholders and digital platforms is essential in shaping a positive digital future.



"Digital spaces must be made safe for those who gather and report the news… When journalists are silenced, we all lose our voice."

UN Secretary-General

unesco

"Since wars begin in the minds of men and women it is in the minds of men and women that the defences of peace must be constructed"

PROTECTING CRITICAL VOICES

Guidance for Human Rights Impact Assessment on Digital Platforms

CONTENTS

PROTECTING CRITICAL VOICES

Guidance for Human Rights Impact Assessment on Digital Platforms

FOREWORD

As providers of information and communication channels for billions of people around the world, digital platforms profoundly impact the rights to freedom of expression and participation, by offering a space for public discourse and civic engagement, creating new forums for public participation and channeling concerns, and amplifying critical voices. Critical voices, particularly journalists and human rights defenders, are vital for promoting public debates, challenging dominant narratives, exposing wrongdoing, advancing justice, accountability, and respect for human rights.

Online spaces and digital platforms have become a central arena for journalists and human rights defenders to perform their roles. At the same time, these are spaces where critical voices are exposed to threats and attacks that may result in censorship or self-censorship, impacting the individuals concerned, but also more broadly impacting societies' access to information and trust in institutions. Attacks faced by journalists and defenders on digital platforms involve unlawful interference with their right to privacy - such as hacking, doxxing, interception, and surveillance, as well as undue limitations to their right to freedom of expression. Many also face direct threats to their life, health and security.[1] Women human rights defenders and journalists, in particular, face heightened risks.

A wide range of digital technologies operate within platforms' ecosystems — from search and discovery services to advertising systems, messaging tools, and generative artificial intelligence (AI) content production. Risks to critical voices can emerge from multiple sources, including the design of these tools or services, as well as from the platforms' own policies and practices.

Online risks are closely tied to the broader offline protection concerns faced by defenders and journalists. Evidence shows that digital attacks are frequently connected to physical violence and intimidation. Risks greatly intensify during critical periods, such as electoral processes and situations of armed conflict. Moreover, attacks by state actors, by those in charge of enforcing the law, represent a major risk factor in certain contexts and are often compounded by legal and extra-legal pressures on companies.

1 OHCHR. 2024. *Civic Space and Tech Brief – Hacking and Spyware.* https://www.ohchr.org/en/documents/tools-and-resources/civic-space-tech-brief-hacking-and-spyware.

1. Aim of this Guidance

This Guidance aims to assist companies in identifying, assessing and responding to risks to human rights associated with digital platforms, particularly in relation to the protection of journalists and human rights defenders.[2]

UNESCO and the Office of the United Nations High Commissioner for Human Rights (OHCHR) jointly developed this Guidance, building on their engagement with a diverse range of stakeholders involved in the protection of civil society actors and proving support and guidance to companies on measures relevant to safeguard human rights in the digital sphere. It rests on the recognition that collaboration between various stakeholders and digital platforms is essential in shaping a positive and responsible digital future, empowering stakeholders to navigate the evolving digital landscape.

For the purposes of this Guidance, a 'digital platform' is understood as providers of software for information-sharing platforms that connects users to facilitate the dissemination of information and content such as social media, search engines, or creative content outlets with relevant presence, size, reach, market share in a specific jurisdiction as outlined in the UNESCO *Guidelines for the Governance of Digital Platforms - Safeguarding freedom of expression and access to information through a multistakeholder approach* (UNESCO Guidelines).[3]

2. What is the framework?

International human rights law provides a universal legal framework that establishes minimum standards for regulating digital platforms.[4] The United Nations *Guiding Principles on Business and Human Rights* (UN Guiding Principles),[5] along with UNESCO's *Guidelines*, the OECD'S guidance,[6] and relevant reports from international human rights mechanisms such as the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression represent key reference points for defining corporate responsibilities. They offer a roadmap for companies' actions and for informing the regulatory initiatives.

As part of their responsibilities, digital platforms are called to conduct human rights due diligence along the value chain and operations to prevent that their services and products have an adverse impact on human rights.[7] Human rights impact assessments (HRIAs), as a component of human rights due diligence, are key processes which help companies to identify, prevent and mitigate potentially negative human rights impacts associated with their products, services and operations. HRIAs specifically identify and evaluate potential human rights risks before they occur.

As set out in the UN Guiding Principles, the process of human rights due diligence includes four core components: identifying and assessing actual or potential adverse human rights impacts that the

---

2 In view of the mandate of UNESCO and OHCHR, the Guidance looks specifically at human rights impact assessments for the protection of critical voices, including journalists and human rights defenders. Nevertheless, most elements in the Guidance would also apply for human rights impact assessments more broadly. For human rights impact assessments related to other specific groups, see for example UNICEF, "Assessing Child Rights Impacts in Relation to the Digital Environment," UNICEF Child Rights and Business, accessed October 16, 2025, https://www.unicef.org/childrightsandbusiness/workstreams/responsible-technology/D-CRIA.

3 UNESCO. 2023. *Guidelines for the governance of digital platforms: safeguarding freedom of expression and access to information through a multi-stakeholder approach.* https://unesdoc.unesco.org/ark:/48223/pf0000387339.

By contrast, the design, development, marketing, sale/licensing and deployment of products, services and solutions will be referred to as the digital platforms' 'own activities', as well as those involving their business relationships, as outlines in OHCHR's B-tech paper. 2020. *Identifying and Assessing Human Rights Risks related to End-Use.* A B-Tech Foundational Paper. www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf.

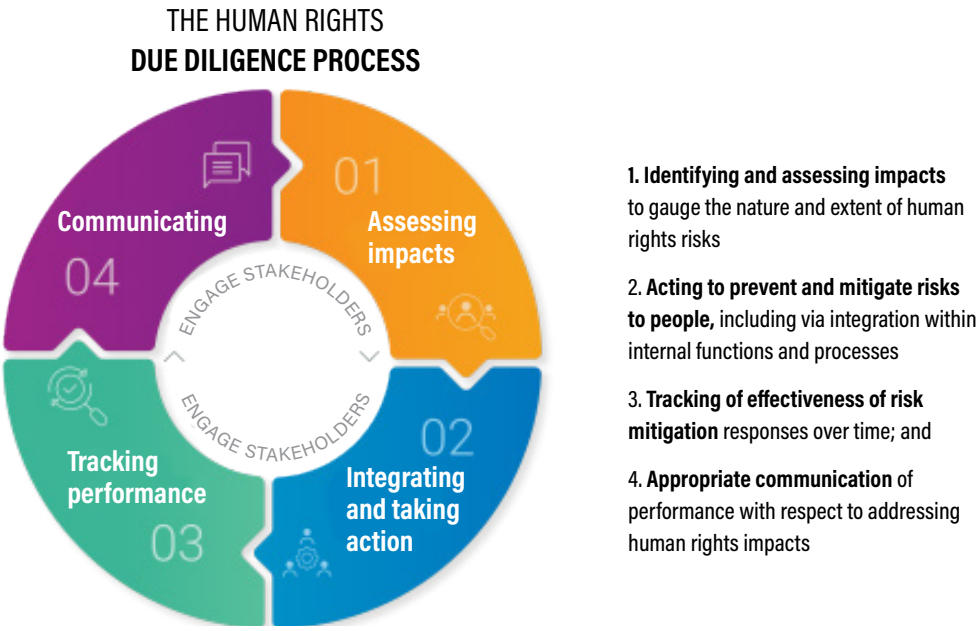
4 The extent of international legal obligations may vary depending on state ratification of the various international human rights treaties. For example, while the International Covenant on Civil and Political Rights (ICCPR) is legally binding for states that have ratified it, the broader guidance of international human rights, contributes to a universal system of human rights.

5 OHCHR. 2012. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework.* https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights.

6 OECD. 2018. Due Diligence Guidance for Responsible Business Conduct. https://www.oecd.org/content/dam/oecd/en/publications/reports/2018/02/oecd-due-diligence-guidance-for-responsible-business-conduct_c669bd57/15f5f4b3-en.pdf; OECD. 2023. Guidelines for Multinational Enterprises on Responsible Business Conduct. https://doi.org/10.1787/81f92357-en.

7 In line with the United Nations Guiding Principles on Business and Human Rights, human rights due diligence refers to the private actors' responsibility to respect human rights by taking adequate measures for their prevention, mitigation and, where appropriate, remediation. Due diligence may also encompass other commitments or activities to support and promote human rights, which may contribute to the enjoyment of rights. Aimed at a wider range of organisations, both for-profit and not-for-profit, the document provides guidance specific to the safety of journalists and human rights defenders.

company may cause, contribute to, or be directly linked to; taking appropriate action and integrating findings from impact assessments across relevant company processes; tracking the effectiveness of measures; and communicating with stakeholders about how impacts are being addressed. While some digital platforms conduct internal risk assessments, often utilizing methodologies and approaches based on the UN *Guiding Principles* and the OECD *Due Diligence Guidance for Responsible Business Conduct*, some states also require companies, including digital platforms, to report on how they manage human rights risks.



Source: OHCHR B-Tech.

The HRIA Guidance responds to the growing recognition of the need to ensure that digital platforms systematically and rigorously assess and mitigate human rights risks faced by critical voices online. HRIAs should improve the understanding of how digital products and services cause or contribute to human rights violations directly or indirectly affecting critical voices and what measures need to be put in place to protect them, while recognizing that these harms can evolve or change over time. In addition, the Guidance helps platforms to align their practices with international human rights standards at the stages of designing and deploying new technologies, as well as during content moderation and curation.[8] The Guidance provides recommendations on measures to make information available to users,[9] establish reporting mechanisms[10] and create specific protection policies for groups in situations of vulnerability and marginalization.[11]

Concerns regarding accessibility and accountability reflected in experiences of human rights defenders and journalists

Digital platforms have played a crucial role in supporting the work of human rights defenders and journalists by enabling research, networking, and information-sharing, particularly in repressive environments. However, these same platforms also impose content restrictions and privacy limitations that hinder their work. Human rights defenders and journalists have faced challenges in accessing remedies and in holding platforms accountable for these restrictions, which impact their ability to communicate and engage with their communities.

8 See Principle 2 of the UNESCO Guidelines, 'Platforms adhere to international human rights standards, including in platform design, content moderation, and content curation.' https://unesdoc.unesco.org/ark:/48223/pf0000387339/PDF/387339eng.pdf.multi.

9 Ibid., Principle 4.

10 Ibid., Principle 5.

11 Ibid., "Context-specific provisions", point 130, paras. a), e), f) and g). https://unesdoc.unesco.org/ark:/48223/pf0000387339/PDF/387339eng.pdf.multi.

A 2023 OHCHR pilot study focusing on the Middle East, North Africa and East Africa,[12] revealed that many human rights defenders and journalists experienced online and offline attacks, including smear campaigns, hacking, phishing, doxxing, and threats of violence, often linked to their online activities. Content removal and account suspensions were common, with many respondents reporting that these actions were taken without clear reasons or effective pathways for appeal. Additionally, they faced issues such as shadow bans and distribution limitations, raising concerns that content moderation policies and algorithms systematically suppress certain types of content. A lack of transparency regarding government requests for content removal further heightened these concerns, particularly for those working in conflict zones.

The pilot study also highlighted the general challenge of dependence on platforms for redress and responses. Human rights defenders and journalists reported significant barriers in addressing their challenges, including limited awareness of platform redress mechanisms, language barriers, and slow or unclear responses to content moderation appeals. Many emphasized the need for platforms to engage more meaningfully with users in shaping content governance policies, especially during critical events like elections and conflicts. They called for improved security measures, clearer moderation policies, and greater transparency in platform operations, particularly regarding government data requests. Without these improvements, trust in digital platforms remains fragile, leaving human rights defenders and journalists at risk of online censorship and threats.

3. Who is this Guidance for?

In accordance with the UN Guiding Principles, the corporate responsibility to respect human rights exists independently of the capacity or willingness of States to discharge their own human rights obligations. This responsibility transcends compliance with national legislation and may necessitate adherence to higher standards in jurisdictions where legal frameworks are inadequate or insufficiently enforced. Consequently, while digital platforms are expected to observe applicable domestic laws, they are concurrently obliged to align their operations with internationally recognized human rights norms, particularly in contexts where State-based protections are deficient.

This Guidance is intended for companies and aims to assist digital platforms in conducting human rights impact assessments, whether or not a state regulatory framework requires such assessments. The Guidance may also provide details that companies can use when reporting to regulatory authorities in jurisdictions where such regulations do exist. It can also serve to inform civil society organizations in their advocacy and engagement with platforms, and guide investors on how to leverage this to promote human rights through corporate decision-making and company policy, and to support auditors by highlighting key elements to consider when assessing digital platforms.

This HRIA Guidance offers a concrete approach to digital platforms for identifying, assessing, mitigating and reporting on human rights risks associated with the design, development, marketing, sale/licensing and deployment of their products, services and solutions that could have adverse human rights impacts on critical voices. The Guidance also serves as a valuable resource for critical voices, those working to protect them, civil society and other actors, such as investors and auditors, in two ways: (i) to engage with companies regarding the needs and risk exposure of critical voices to prevent harm; and (ii) to hold platforms accountable. It provides an overview of recommendations for mitigating threats and minimizing risks and could inform specific training activities or funding initiatives aimed at empowering those working to protect human rights in digital spaces.

12 See www.ohchr.org/sites/default/files/documents/issues/civicspace/Results-overview-of-pilot-study-on-experiences-with-social-media-and-communication-platforms-in-MENA-and-East-Africa-regions-June-2023.pdf.

DEVELOPMENT OF THE HUMAN RIGHTS IMPACT ASSESSMENT GUIDANCE (HRIA)

In 2022, on the occasion of the 10th anniversary of the United Nations Plan of Action on the Safety of Journalists and the Issue of Impunity[13] and in relation to its strengthened implementation, UNESCO convened a multi-stakeholder consultation in Copenhagen, Denmark, for Human Rights Impact Assessment Guidance on the Safety of Journalists Online, resulting in the development of the first version of this Guidance.

This Guidance was reviewed and updated based on the outcomes of several global consultations,[14] which eventually led to the UNESCO Guidelines, which were finalized and published in November 2023. Additionally, it drew on insights from a series of consultations organized by UNESCO in 2024 and 2025 across various international forums.[15] These consultations brought together international regulatory authorities,[16] media councils, journalists, media organizations, civil society, and academia from various regions, drawing on their experiences and insights. This aimed to ensure that the Guidance was aligned with the realities faced by diverse stakeholders and that it incorporated their suggestions for improving comprehensive risk evaluation and prioritization.

In September 2024, States adopted the UN Global Digital Compact and with it, OHCHR's human rights advisory service on digital technologies. Through the pilot phase of the advisory service including consulting, commenting and engaging in content governance laws all over the world OHCHR developed a way forward on 'Online Platform Governance and Human Rights'[17] The consultative processes in relation to national laws and practices have also informed this Guidance.

Furthermore, the Guidance is aligned with the objective of the Global Digital Compact to foster an inclusive, open, safe and secure digital space,[18] through upholding human rights and ensuring accountability in the digital sphere. It reflects the emphasis of the Compact on integrating human rights due diligence and impact assessments throughout the technology lifecycle, as well as promoting accountability mechanisms to prevent abuses and ensure access to effective remedies. By embedding these principles, the Guidance reinforces the shared goal of creating a digital ecosystem that respects and protects human rights while advancing global development objectives.

This Guidance builds on and promotes consistency with the guidance developed by the B-Tech Project[19] of OHCHR and supports digital platforms in the implementation of resolutions on freedom of expression, safety of journalists and human rights defenders at the United Nations General Assembly and Human Rights Council, as well as the United Nations Plan of Action on the Safety of Journalists and the Issue of Impunity, and other UN policies and action plans.[20]

---

13 UNESCO, "UN Plan of Action on the Safety of Journalists and the Issue of Impunity," accessed October 16, 2025, https://www.unesco.org/en/safety-journalists/un-plan-action.

14 https://www.unesco.org/en/internet-trust/guidelines-consultation-process?hub=71542

15 The 11th edition of the Forum on Internet Freedom in Africa (FIFAfrica), held in September 2024, in Dakar, Senegal, see https://cipesa.org/event/fifafrica24/; the Latin American Conference on Investigative Journalism, held in October 2024, in Madrid, Spain, see https://colpin.ipys.org/; the 20th African Investigative Journalism Conference, held from 30 October–1 November 2024, in Johannesburg, South Africa, see https://aijc.africa/; the International Day to End Impunity of Crimes Against Journalists (IDEI), in November 2024, in Addis Ababa, Ethiopia, see https://unesdoc.unesco.org/ark:/48223/pf0000390976; World Press Freedom Day 2025 in Brussels, Belgium, see https://www.unesco.org/en/days/press-freedom#:~:text=Every%20year%2C%203%20May%20is,the%20exercise%20of%20their%20profession.

16 See Global Forum of Networks of Regulatory Authorities. https://www.unesco.org/en/internet-trust/building-network-networks?hub=71542.

17 Office of the United Nations High Commissioner for Human Rights (OHCHR), Civic Space and Online Platform Governance: Briefing Note (Geneva: OHCHR, 2022), https://www.ohchr.org/sites/default/files/documents/issues/civicspace/resources/civic-space-online-platform-governance-brief.pdf.

18 See Objective 3 of the Global Digital Compact "Foster an inclusive, open, safe and secure digital space that respects, protects and promotes human rights", and specifically §22, §25 a) and b).

19 The OHCHR B-Tech Project provides authoritative guidance and resources for implementing the United Nations Guiding Principles on Business and Human rights. https://www.ohchr.org/sites/default/files/Documents/Issues/CivicSpace/UN_Guidance_Note.pdf.

20 See United Nations. 2020. *Guidance Note of the Secretary-General: Protection and Promotion of Civic Space.* www.ohchr.org/sites/default/files/Documents/Issues/CivicSpace/UN_Guidance_Note.pdf and United Nations. 2012. *Plan of Action on the Safety of Journalists and the Issue of Impunity.* https://www.ohchr.org/sites/default/files/documents/issues/journalists/2023-01-31/un-plan-on-safety-journalists_en.pdf.

ADDRESSING HUMAN RIGHTS IMPACTS

The steps in this HRIA Guidance encompass:

1. Participatory context and local stakeholder analysis.
2. Identification of risks and impact on human rights
3. Mitigation measures.
4. Review and update.

> **ELEMENTS TO ASSESS HUMAN RIGHTS IMPACTS** (THE POTENTIAL OF EVENTS TO CAUSE HARM TO HUMAN RIGHTS):
>
> - **THREAT:** any indication or declaration of intent to inflict harm to human rights, whether recent or imminent.
> - **VULNERABILITY:** any factor within the digital service that increases the likelihood or severity of adverse impact.
> - **CAPACITY:** any resource and abilities that enhance security, control and correct risk.
>
> The interplay of these factors can either increase (through threats and vulnerabilities) or decrease (through capacities) the overall human rights impact. Consequently, to effectively manage risks, it is essential to address threats and vulnerabilities while also bolstering capacities.

1. Context analysis: Understanding the environment in which risks may manifest[21]

Digital platforms operate within complex and rapidly changing environments. The first step for developing a human rights impact assessment is understanding the regional and local contexts, considering the unique specificities of each area where a digital platform operates, recognizing that the challenges faced by human rights defenders and journalists may be vastly different depending on the region, country and local context in which they work, as well as the type(s) of product and features that each platform provides. This requires a thorough analysis of the stakeholders, and an understanding that risks can fluctuate significantly based on evolving threats, vulnerabilities, and capacities. Therefore, a situational analysis should be continuously updated in consultation with relevant stakeholders – especially diverse local voices – to evaluate the likelihood and degree of impact and to define the measures that should be taken to mitigate adverse impact. By being implemented with relevant stakeholders on the ground, this context analysis will not only be informed by the perspective of digital platforms but also depend on or benefit from the local expertise and a comprehensive overview of the situation.

Specifically, the digital platforms should:

Engage with the stakeholders who are likely to be most affected to adapt their internal control and risk management policies. To conduct a comprehensive context analysis, digital platforms should engage with various stakeholders, including local authorities, policymakers, regulatory bodies, as well as civil society, and human rights defenders and journalists themselves (including women, LGBTQIA+ persons, people with disabilities and Indigenous Peoples). Such engagement offers

21 As outlined later in the document, contextual analysis might include the legal and regulatory environment, economic and technological infrastructure, social and cultural dynamics, political and security conditions, and the specific risks faced by marginalized groups and voices.

practical insights into the digital environment of their country or region, providing a well-rounded understanding of the risks and challenges faced by human rights defenders and journalists. Through this engagement digital platforms should at a minimum seek to:

- **Map the patterns and violence against critical voices** in the digital environment, including whether such violence is prevalent around certain events like elections or protests or if it targets critical voices with specific protected characteristics (women, LGBTQIA+ communities, Indigenous Peoples, people with disabilities).
- **Assess the frequency and nature of online attacks and risks**, including those resulting from innocuous and passive consequences and system design.
- **Assess the frequency and nature of online attacks** that have translated into physical violence and other offline repressive acts against critical voices.
- **Determine the level of impunity** for attacks against critical voices, which can affect the likelihood of such incidents recurring on digital platforms.
- **Investigate if certain critical voices are disproportionately targeted** due to their work on specific topics.
- **Understand if the type of online attacks faced by local or community-based critical voices** are increasing or different from those experienced by well-known or international critical voices.
- **Evaluate whether women, LGBTQIA+ persons, Indigenous peoples, critical voices or persons with disabilities**, face comparatively higher or different risks.
- **Consider linguistic diversities**[22] to better understand the specific realities of the countries or regions where they operate. This requires the ability to function in multiple languages, including those that are underrepresented on the internet.
- **Understanding the political, social, legal and economic contexts of the regions where they operate,** including crisis situations.

Assess the state of rule of law and related challenges, and then adapt internal control and risk management policies without compromising responsibilities under international human rights law: This includes aspects such as supremacy of the law, compliance with the law, law-making procedures, degree of legal certainty, prevention of abuse of powers, equality before the law and non-discrimination, access to justice, independence and impartiality of the judiciary, access to courts and fair trial, effectiveness of judicial decisions, corruption, repressive measures, surveillance practices, etc. Such insight enables platforms to better assess the environments in which they operate and make more informed and appropriate decisions, particularly when it comes to their potential cooperation with governments and responses to government requests. Understanding how these conditions intersect with the challenges faced by critical voices in the digital environment is crucial for assessing the unique risks they encounter and for devising targeted protection strategies. Platforms should always anchor internal systems and processes on international human rights law, including when the local legal framework does not align with them.

Evaluate the legal and policy framework and practice of the country/region: Digital platforms should know and understand the legal and policy framework and practice of the country/region where they operate, while keeping in mind that these considerations are not static. This understanding enables platforms to:

22

- **Assess the role of digital platforms** as significant or dominant means of content dissemination and data collection.
- **Identify and assess potential sources of threats from various actors, such as:**
  - Government and political leaders who have a history of hostility or general animosity towards critical voices.
  - Non-state actors, including organized crime groups, that threaten critical voices to prevent exposure of their activities.
  - Private corporations that may attempt to suppress critical reporting or influence media narratives.
  - The general public, which can engage in harassment or intimidation campaigns against critical voices through social media or other platforms, often in coordinated manners.[23]
- **Evaluate the regulatory framework**, existing legal restrictions and human rights and protection mechanisms to assess their alignment with the human rights standards.
- **Identify (if any) relevant regulatory authorities** and understand their role, independence and influence in the region or country.

Assess how emergencies and crisis situations, such as armed conflicts, national unrest, and natural disasters may affect their work and influence their ability, including those in exile, to operate safely and effectively.

Addressing structural risks related to legal and institutional frameworks and practices

Effective risk management strategies should be context-specific, requiring a clear understanding of how legal and institutional frameworks affect public freedoms and, particularly, how authoritarian practices and the degree of rule of law shape society – especially civil society, at different levels. It also requires critical consideration of how such contextual risks may be exacerbated or enabled due to risks that are systemic to the design and operation of the services themselves.

Criminalization of dissent, lack of minimal independent judicial oversight, and weak accountability for law enforcement abuses often result in systemic threats to critical voices. To mitigate these risks effectively, digital platforms should establish meaningful and safe communication channels and accessible redress mechanisms. Given that many individuals at risk operate in remote or resource-limited areas, some affected by conflict or extreme public insecurity, channels to these actors must be both easy to access and specifically equipped to process urgent demands.

In different contexts, deep mistrust exists between civil society and government authorities. As a result, confidence in corporate risk mitigation strategies depends heavily on effective transparency in terms of how corporations engage with authorities. Digital platforms should disclose how they respond to legal and extra-legal requests from authorities, as well as provide information regarding their efforts to challenge demands that are inconsistent with human rights.

23 See the UNESCO. 2024. *Guide for Journalists Covering Hate Speech: A Guide for* Journalists. https://unesdoc.unesco.org/ark:/48223/pf0000392378. 'Hate campaigns usually try to goad ordinary citizens into action and show that public opinion is on their side. The demand side of hate speech is complex but often has to do with underlying insecurities about economic conditions or rapid social change that have little to do with the identity group being targeted for hate.'

DIFFERENTIATED IMPLEMENTATION

Human rights impact assessments must be implemented with a differentiated approach. This involves enhancing the ability of digital platforms to identify adverse impacts among users and knowing how to adapt their actions based on the specificities identified.[24] The differentiated approach requires understanding that risks and impacts may differ depending on the region and context that human rights defenders and journalists operate in, necessitating approaches that are tailored to their protection needs as opposed to a one-size fits all approach.

This differentiated approach should be transversal to the entire human rights impact assessment, from the time of capturing the information and recording it, in the interaction with critical voices, in the analysis of the context, in defining the likelihood and severity of the adverse impact, when deciding the measures and in the reporting, reviewing and updating process. In other words, the application of a differentiated approach goes far beyond the inclusion of some elements in the risk assessment process.

It is not just about understanding the specific challenges to journalists and human rights defenders in various regions, but about safeguarding, empowering, and protecting the work they do, and to adapt mitigation measures in response to risks they face on platforms.

This may include adopting mitigation measures for critical voices who belong to one or several categories or groups that are at higher risk or who face particular forms of risks that require an intersectional approach to ensure effective protection. This includes women and girls, migrants, Indigenous peoples, people belonging to ethnic or religious minorities/groups, people deprived of liberty, LGBTQIA+ persons, older adults, people with disabilities, and children, among other relevant groups. But it also requires adopting and adapting mitigation measures that respond to local realities which may be widely different.

Stakeholder analysis and engagement: evaluating individuals and groups at risk

Digital platforms should conduct a thorough stakeholder analysis to identify and evaluate the human rights impacts of their products and services on specific individuals and groups, such as human rights defenders and journalists, including those in heightened situations of risk. This also entails engaging with affected stakeholders, also by adjusting channels for inputs and feedback to their needs and capacities. Understanding their unique risks and involving them in identifying pathways towards minimizing them to help tailor appropriate responses. This also implies considering factors such as their role, geographical location, and any specific challenges they might face due to their identity or situation.

For instance, Indigenous defenders and journalists may face unique challenges and risks online compared to their peers due to cultural, geographical, or socio-political factors.[25] Such specific issues should be addressed to assess adverse impacts adequately.

---

24 The OHCHR interpretative guidance for the United Nations Guiding Principles outlines the following: 'Depending on the operational context, the most severe human rights impact may be faced by persons belonging to groups that are at higher risk of vulnerability or marginalization, such as children, women, Indigenous Peoples, or people belonging to ethnic or other minorities. If the company decides it needs to prioritize its responses to human rights impacts, it should consider the vulnerability of such groups and the risk that a delayed response to certain impacts could affect them disproportionately.'

25 For example, if online violence tends to target journalists in general, online violence against women show some specific characteristics:
- It is networked;
- It is usually misogynistic;
- It radiates;
- It is intimate;
- It can be extreme, intense and prolific; and

See p. 11 of Posetti, J., Shabbir, N., Maynard, D., Bontcheva, K., and Aboulez, N. UNESCO. 2021. *The Chilling: global trends in online violence against women journalists*; research discussion paper. https://unesdoc.unesco.org/ark:/48223/pf0000377223.

> Therefore, incorporating intersectionality into human rights impact assessment processes is crucial. The safety of critical voices can be compromised not only due to the nature of their work but also due to intersecting factors like gender, age, race, and other forms of discrimination.
>
> By addressing these intersecting factors, digital platforms can develop more effective human rights impact assessment strategies that reflect the complex realities faced by critical voices in their interactions with their products and services.

GENDER-RESPONSIVE

A gender-responsive perspective should be integrated into all analyses of risks and inequalities, not just those specifically concerning women human rights defenders and journalists. This approach helps to address broader issues related to traditional roles and power dynamics.

CRISIS RESPONSIVE

HRIAs should serve as the foundation upon which crisis protocols are built. By analyzing the likelihood and severity of human rights impact for critical voices in crisis contexts, digital platforms can categorize areas of concern and tailor mitigation measures to each category. This proactive approach allows for informed decision-making and targeted mitigation strategies before violations occur, aligning with the preventative spirit of human rights due diligence frameworks as provided by the UN Guiding Principles. For example, digital platforms should identify and assess all risks stemming from their products, services and operations which are deployed in a crisis situation or high-risk areas and that may contribute to escalating a crisis.

Crisis protocols should be developed in direct response to the specific risks identified through the HRIA process. These protocols detail the actions a platform will take when a crisis situation emerges, and tailor measures based on the type of crisis, which can consist of different scenarios such as the rapid spread of hate speech during political unrest or elections, armed conflict, natural catastrophes, internet shutdowns, cyber-attacks or other forms of crisis. Without an initial risk assessment, crisis responses may end up being generic, reactionary, or misaligned with the actual vulnerabilities of the platform and its users. Embedding crisis planning within the broader framework of impact assessments ensures that response measures are both relevant and robust, ultimately strengthening the platform's ability to uphold human rights under pressure.

Crisis protocols should be context-specific, developed in collaboration with local experts, and regularly reviewed to ensure they remain responsive to evolving threats, as needs and urgency can fluctuate during different phases of a crisis or conflict. A robust crisis protocol should go beyond risk identification and the severity test, and include clear procedures for escalation, communication, and mitigation. Tools such as a risk matrix may support this process.[26] Crisis protocols need to be accompanied by a dedicated team, with expertise in crisis management and that can involve people with deep knowledge of local contexts, including linguistic, political, social, cultural and legal aspects.

2. Identification of risks and impact on human rights

Critical voices face a range of human rights risks when using digital services and products, affecting their rights to liberty and security, freedom of expression and access to information, privacy, equality and non-discrimination among other human rights. These include acts such as smear or defamatory campaigns, direct threats of violence, account hacking or phishing, false impersonation, doxxing on their social media, and more.

---

26 See International Media Support (IMS). 2025. *Guide For Risk Management in the Context of Emergencies, Armed Conflicts and Crises – Based on contextual analysis of the situation in Ukraine*. https://www.mediasupport.org/publication/guide-for-risk-management-in-the-context-of-emergencies-armed-conflict-and-crises/.

Inherent risks related to data and privacy

Digital platforms present significant privacy risks due to their collection and storage of vast amounts of personal data, including user behaviour, search history, contacts, and location. Weak or inadequate privacy policies, sometimes combined with inadequate or non-existent legal protections for data and privacy by the State can lead to data exploitation, security breaches, and unauthorized access by third parties including malicious or abusive state actors through government requests.

Importantly, the problem with digital platforms privacy policies is not merely their inadequacy, but their structural flaws: they are typically non-negotiable, offering users little to no meaningful control over how their data is processed. They are rooted in business models in which data extraction and monetization are core to the platform's revenue, often regardless of whether a user pays for the service. A user's right to privacy and informed consent over their data should not depend on whether they pay for a service.

It is therefore essential to assess these impacts comprehensively, considering both platform infrastructure and operations, as well as contextual factors such as national legislation, the human rights track record and the state of rule of law, ongoing conflicts, and sociopolitical divides. Human rights impact assessments should be supported by robust privacy safeguards, independent oversight mechanisms, and user empowerment measures to help individuals retain greater control over their personal data.

IDENTIFICATION OF RISKS

The below table is not an exhaustive list of human rights risks but offers instead examples to help digital platforms identify potential human rights risks in design, development, marketing, sale/licensing, use and misuse of products, services and solutions that can adversely impact the human rights of critical voices.

HUMAN RIGHTS RISKS TO CRITICAL VOICES ONLINE	WHAT IS IT AND HOW DOES IT IMPACT HUMAN RIGHTS?	EXAMPLES OF FEATURES OF PRODUCTS; SERVICES; AND POLICIES
HARASSMENT AND ONLINE THREATS	One of the most prevalent risks is online harassment, including threats, defamation, and trolling campaigns. This can take the form of verbal abuse, hate speech, or organized efforts to discredit and intimidate critical voices. Women critical voices, in particular, face heightened risks of gender-based harassment, including sexualized threats and attacks aimed at undermining their credibility. This abuse can lead to psychological distress, self-censorship, and, in extreme cases, is combined with physical threats (psychological and physical harm) ultimately silencing them, undermining everyone's access to information and directly curtailing their freedom of expression and their right to security. Similarly, content produced by critical voices can be flagged or removed on the orders of governments, and critical voices' accounts can be blocked, leading to increased censorship. Such practices may be particularly prevalent in crisis situations, such as during armed conflict or internal unrest. Harassment and online abuse can escalate into offline violence, particularly in contexts where the rule of law is compromised and in situations of armed conflict or extreme public insecurity.	■ Live streaming and real-time features ■ Anonymity and encryption features ■ Identity verification systems ■ Commenting and reply systems ■ Tagging, mentioning and sharing ■ Localization features ■ AI-generated content features ■ Content moderation rules ■ Advertising models ■ Redress schemes

DISINFORMATION AND SMEAR CAMPAIGNS	Critical voices may be the targets of coordinated smear campaigns aimed at discrediting their work and/or reputation. These campaigns are often fuelled by false information and aim to undermine trust in critical voices and civil society organizations and media outlets they are affiliated with. Disinformation and smear campaigns lead to reputational harm, reduce public confidence in human rights activism and journalism, and in grave cases, culminate in offline attacks.	■ Algorithmic content curation and moderation ■ Content moderation rules ■ AI-generated content features ■ Sock puppet accounts ■ Crowd-sourced fact-checking features ■ Identity verification ■ Redress ■ Advertising model ■ Fact-checking and labelling policies
DOXXING AND OTHER PRIVACY VIOLATIONS	Doxxing, or the disclosure of personally identifiable information (e.g. home address, phone numbers) without consent, can be used to intimidate and silence critical voices. This tactic endangers not only the individuals at risk, but also their loved ones. In addition, doxxing is often followed by threats, stalking and other forms of harassment. Furthermore, the disclosure of location or personal data can lead to physical threats, especially for people in conflict zones, repressive regimes or high-risk environments. Doxxing often takes a gendered and discriminatory slant: women are often the targets of such campaigns, with sexualized content posted online without their consent as a form of threat, retaliation, or blackmail.	■ Personally identifiable information features such as tagging, mentioning, facial recognition, as well as localization features ■ Data access and collection features such as availability of followers list and connections ■ Content sharing and amplification features ■ Security features ■ Encryptions features ■ Content moderation policies (doxxing and harassment policies); user reporting and redress ■ Account verification ■ Data protection policies
UNLAWFUL SURVEILLANCE	The data-driven nature of digital platforms allows for the collection and sharing of personal data and other information that may place critical voices at particular risk. As a result, the data of HRDs, journalists and others can become the target of unlawful surveillance via the platforms. These intrusive hacking tools can compromise their communication, capturing their contacts and the sensitive information they collect through their work. Spyware technology undermines the protection of critical voices and also exposes their contacts and sources, putting them at significant risk.	■ Invasive application permissions ■ Cross platforms tracking ■ Location features ■ Encryption features ■ Metadata storage ■ Cloud security features ■ Anti-tracking protections ■ HTTPS/TLS implementation ■ Multi-factor authentication (MFA) ■ Data retention ■ Privacy and data protection policies ■ Attention to government requests policies ■ Security and authentication policies ■ Encryption policies ■ Anonymity protections

ALGORITHMIC BIAS AND PLATFORM CENSORSHIP	Automated moderation systems implemented by digital platforms can lead to the suppression or demotion of non-harmful and legitimate content produced by critical voices (e.g. politically sensitive content). Algorithms can also contribute to the amplification of harmful narratives against these voices, amplifying the harassment and abuse against them. Moreover, there are documented examples of the platforms' advertising systems inadvertently promoting hate speech or incitement to hostility or violence. There is an emerging concern regarding the involuntary or unauthorized access to sensitive data on digital platforms by cloud service providers. This encompasses scenarios where cloud companies might use stored data to train AI models without explicit consent, share such data with authorities without proper safeguards, or even restrict access to essential cloud services, undermining the ability of these voices to operate securely and freely.	■ Algorithmic content curation and moderation ■ AI-generated content features ■ Automatic monetization features ■ Fact-checking labels ■ Content moderation and curation policies ■ Hate speech and disinformation policies ■ Human oversight and moderation policies ■ Algorithmic transparency ■ Advertising policies

IDENTIFICATION OF THE LIKELIHOOD AND THE SEVERITY OF ADVERSE HUMAN RIGHTS IMPACTS

Once potential risks have been identified, digital platforms should focus on assessing the likelihood and severity of adverse impacts of the risks. This assessment involves analyzing how different threats and vulnerabilities in the products and services match existent capacities or require additional capabilities to prevent or mitigate risks. The table below highlights one example of many risks.[27]

TABLE 1. EXAMPLE OF RISK ASSESSMENT FACTORS

RISK	THREAT	VULNERABILITY	CAPACITIES (EXISTENT)
SMEAR CAMPAIGN AND REPUTATIONAL DAMAGE	Hacking of accounts and digital impersonation	Depends on the characteristics of the digital service (some examples): ■ Weak authentication mechanisms ■ Phishing and social engineering susceptibility ■ Application programming interfaces (APIs) and integration weaknesses ■ Inadequate identity verification	Depends on the characteristics of the digital service (some examples): ■ MFA ■ Strong password policies ■ Secure APIs ■ Regular security audits ■ Channels for companies to take down impersonation accounts (provided they are not parody accounts)

Once the risk factors for critical voices within digital platforms are analyzed, companies should determine the likelihood and severity of impact of these risks and level them as high, medium and low.

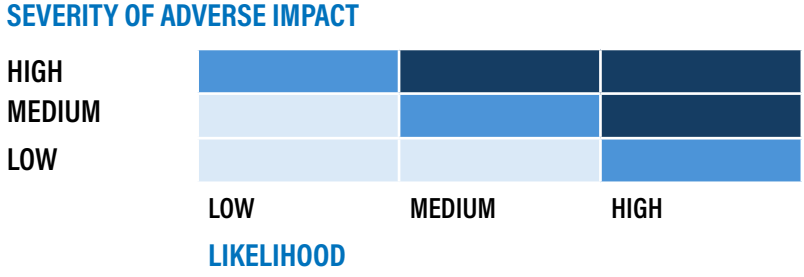
The severity analysis helps to prioritize mitigatory action. The likelihood and impact of a risk depend on diverse elements such as the **scope** (how many people could be affected by the risk?), **scale** (how severe would the harm be?), its **mitigation or remediability** (can a remedy fully restore the harm caused?), its past occurrence, among other factors that could raise depending on the context analysis. For example, some disinformation risks against critical voices increased their likelihood

27 Election periods can indeed represent a great risk for internet users, and this is recognized through initiatives led by big tech companies, such as the 2024 *Tech Accord to Combat Deceptive Use of AI in 2024 Elections* published in February 2024, where tech companies set expectations on how they will manage the risks arising from deceptive AI election content. Signatories to the Tech Accord include Adobe, Amazon, Anthropic, Arm, ElevenLabs, Google, IBM, Inflection AI, LinkedIn, McAfee, Meta, Microsoft, Nota, OpenAI, Snap, Stability AI, TikTok, TrendMicro, Truepic, and X.

and severity of impact during electoral processes, protests, emergencies, government crackdowns, cyberattacks, crises such as armed conflicts and humanitarian crisis. Risks are influenced by the effectiveness of mitigation measures.

Digital platforms should be able to review the likelihood and severity of impact for each risk of harm. The guiding principle when deciding what contextual elements to collect should be whether it will improve the accuracy of the risk assessment and the understanding of harm.

Moreover, prioritization of mitigatory action would also be informed by the severity of the adverse impact. Meaning that digital platforms should prioritize action when they identify risks whose scale, scope and remediability shows that it will seriously affect the human rights of critical voices.[28]



Depending on the risk level, digital platforms would need to decide on the measures that are appropriate to reduce the risk of harm to critical voices.

Digital platforms need to communicate with the stakeholders engaged in the outcomes of the risk assessment (unless this increases the risks).

3. Mitigation plan

Critical voices such as journalists and human rights defenders have a unique relationship with digital platforms, driven by the nature of their work, which centres on the constant flow of information – whether they are seeking, receiving, or publishing it. For these professionals, digital platforms are indispensable tools for disseminating information to their audiences, engaging with victims/survivors and sources, verifying facts, and staying updated on global events in real time.

Any measures implemented by digital platforms to ensure the safety of critical voices should extend beyond personal protection; they should also ensure the security of the information they handle.

Mitigation encompasses both preventive and responsive measures. Preventive measures involve adopting approaches to reduce the likelihood of risks occurring in the first place. Responsive measures, on the other hand, focus on managing and minimizing the impact after a risk has materialized.

PREVENTIVE MEASURES

At the minimum, digital platforms should ensure they have the systems and processes in place to mitigate risk by implementing, among others, the following **preventive** measures.[29]

28 In this regard, the OSCE has published Guidelines for monitoring online violence against female journalists which include 15 risk indicators for violence escalation. See OSCE, Posetti, J, Maynard, D. Shabbir, N. 2023. *Guidelines for Monitoring Online Violence Against Female Journalists*. https://www.osce.org/files/f/documents/b/0/554098_1.pdf.

29 See Search for Common Ground, Integrity Institute, and Council on Technology & Social Cohesion. 2025. *Prevention by Design: A Roadmap for Tackling Technology-Facilitated Gender-Based Violence at the Source*. 2025. See also UltraViolet. 2021. *New Report Card Grades Social Media Platforms on Handling of Harassment, Hate Speech, Misogyny, Disinformation*; and reports from #ShePersisted. https://she-persisted.org/our-work/research-and-thought-leadership/.

1. Enhanced privacy, security and monitoring features

a. **End-to-end encryption**: Some digital services like messaging applications should mandate end-to-end encryption for communications to ensure that only the intended recipients can read the content. This protects sensitive information and conversations from being intercepted, along with journalistic sources.[30]

b. **Two-factor authentication (2FA)**: Encourage and require the use of multiple options of 2FA (text and SMS-based 2FA, authenticator applications, hardware tokens, etc.) to add an extra layer of security to critical voices' accounts.

c. **Anonymity tools**: Provide tools that allow critical voices to operate anonymously if needed, including secure browsing modes, pseudonym use, and masking internet protocol (IP) addresses.

d. **Threat monitoring and mitigation tools**: Platforms should include systems to detect and flag online harassment, doxxing attempts, deepfake content, or breaches targeting critical voices. These tools should provide them with real-time alerts and steps to mitigate risks.

e. **Protection against unlawful surveillance**: Platforms should strengthen security protocols to protect against unlawful surveillance tools, such as spyware and other forms of unlawful surveillance of files and communications shared on digital platforms.

f. **Default privacy settings to minimize user vulnerability:** Platforms should automatically set strong privacy defaults for users, especially critical voices, to minimize their exposure to risks, with the option to adjust settings for more customization if needed. This ensures that users don't have to take extra steps to protect their privacy and can focus on their work without the additional burden of constantly managing security risks.

   Platforms should also consider:

   - **Enhanced privacy systems**: Advanced privacy controls, such as customizable settings, should allow users to determine who can view their content, follow them, and interact with them, while selective sharing options enable posts to be shared with specific groups or individuals. Such features may be particularly important for critical voices at heightened risk of online harassment, including women and LGBTQIA+.
   - **Profile anonymity tools** that can help critical voices use pseudonyms or conceal identifying details without compromising their professional reach.
   - **Filters in empowering users in managing content exposure**: Platforms should provide users with customizable filters to limit exposure to harmful or unwanted content such as graphic imagery, triggers and specific keywords. Filters can support the mental well-being of critical voices online.
   - **Red teaming:** Deliberate testing of a system's vulnerabilities to identify security risks and identify vulnerabilities, such as those involving privacy, in order to make targeted improvements.[31]

2. Collaboration with civil society, including at local levels

a. **Collaboration with civil society organizations (CSOs):** Collaborate with CSOs that specialize in protecting critical voices acting globally and locally and particularly with organizations representing groups targeted by violence and discrimination. Platforms should work with these CSOs to develop and tailor safety protocols, provide support in cases of digital harassment, establish or bolster escalation channels.[32] These collaborations could involve consultations with a range of stakeholders to ensure their concrete experiences are incorporated in the assessment process ensuring that the measures designed are effective.[33]

---

30 OHCHR, *Civic Space & Tech Brief: Encryption*. 2024. https://www.ohchr.org/en/documents/tools-and-resources/civic-space-tech-brief-encryption-0.

31 See UNESCO, Chowdhury, R., Skeadas, T., Lakshmi, D., and Amos, S. 2025. *Red Teaming artificial intelligence for social good - The PLAYBOOK*. https://unesdoc.unesco.org/ark:/48223/pf0000394338.locale=en.

32 A crisis mechanism that CSOs use to communicate directly with the trust and safety teams within a digital platform to facilitate more attentive, effective support on behalf of individual users who are particularly at risk, like journalists or human rights defenders.

33 See PEN America. 2024. *The Power of Peer Support: Helping Journalists Persevere in the Face of Online Abuse*. https://pen.org/report/peer-support/.

b. **Providing support and data:** Encourage and support independent observatories and initiatives to access data, monitor, and address coordinated and automated harassment campaigns. For example, no content distributor has all the answers to stopping technology-facilitated gender-based violence (TFGBV). Acknowledging this and collaborating with independent and specialized observers and groups which research such harms by allowing safe access to data and trends that will allow all parties to better understand such harms, and work towards finding innovative solutions.

3. Transparency and accountability[34]

a. **Meaningful transparency reports:**[35] Transparency reports are part of how digital platforms assess and report on the scope of their impact. While many platforms often publish numbers, such as how many pieces of content are removed, these figures do not always reflect the actual human rights impacts. Digital platforms should regularly publish meaningful transparency reports detailing how platforms are addressing threats against critical voices, including data on content removal and user bans related to harassment or threats, requests from law enforcement and other government requests. At a minimum, meaningful transparency requires the disclosure of: (i) the number and type of complaints received, how these complaints are being processed and specification of the status of these requests; (ii) whether changes are made to the design, policy or practice of the platforms; (iii) the number and type of government requests to block/remove content and critical voices user' accounts, including the legal basis of such requests, the type of content concerned, and how the platform responded to the requests; (iv) the number and type of government requests for data, and how the platform responded to the requests. To enhance accountability and usability, the reports should be made comparable by adhering to certain standards (e.g. machine readability) and should include disaggregated data to allow for more granular analysis.

Communicating transparently

The UNESCO Guidelines state that digital platforms should regularly report to the public and the governance system on how they adhere to the principles of transparency and explicability, and how they perform relative to their terms of services and community standards. Transparency should be meaningful – the information provided should be as clear and concise as possible, sufficiently detailed and as complex as necessary.[36]

In this regard, the UNESCO Guidelines outline that in any kind of regulatory arrangement, digital platforms should be able to demonstrate the systems or processes they have established to ensure ongoing human rights due diligence, including human rights and gender impact assessments, as well as risk mitigation measures.

In the same tenure, as highlighted in the UN Guiding Principles, companies are expected to be transparent about how they are addressing human rights risks, which includes sharing progress and challenges.[37]

The Guiding Principles establish that companies should 'account for how they address their human rights impacts' and 'be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders' (UN Guiding Principle 21). They also state that 'communication can take a variety of forms, including in-person meetings, online dialogues, consultation with affected stakeholders, and formal public reports.' The practical implications of these considerations in the context of the use of a technology product or service would be useful to explore and elaborate through multi-stakeholder processes.

34 It is important to consider complying with all the transparency recommendations set on the UNESCO Guidelines, paras. 111–118.
35 See transparency reports of the Digital Services Act. 2022. https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency.
36 See paras. 111 and 112 of the UNESCO Guidelines. https://unesdoc.unesco.org/ark:/48223/pf0000387339.
37 OHCHR B-Tech Foundational Paper. 2023.*Key Characteristics of Business Respect for Human Rights*. www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.

A few things are clear:

– In light of the often-large knowledge and expertise gap between technology company personnel and the public about how many technology products, services and solutions work, effective communication will be a critical part of engendering trust with users, customers, society-at-large and policymakers.

– Communication and reporting should include a focus on how impacts have been identified and what impacts have been identified. But it should also involve transparency about the mitigation steps technology companies are taking to address impacts (whether in their own activities or via the use of leverage) and include an evaluation of the effectiveness of those steps. This will involve going beyond the valuable step of publishing findings from impact assessment processes.

– Certain considerations will be particularly important when companies are communicating about end-use scenarios, including:

a) Not putting affected stakeholders or users at risk or in any way undermining their rights to privacy.

b) Not broadcasting technological solutions in ways that allow ill-meaning actors to then combat prevention and mitigation steps.

c) Ensuring that legitimate concerns about commercial sensitivity do not undermine the importance of accountability, transparency and shared learning.

b. **Access to third-party controls:** Enable access to third-party controls to foster innovation in user protection and increase independent oversight. As transparency could also entail external audits of risk assessments, the outcomes of risk assessment processes, mitigation measures, should be submitted to external reviewers.

c. **Shifting away from engagement-driven content ranking**: Engagement-based algorithms often prioritize content that trigger emotional reactions, such as anger and fear, and consequently amplify harmful narratives, polarization and targeted attacks, including against critical voices. Platforms should shift toward content-ranking systems that prioritize trustworthy, contextually relevant, and rights-respecting information.

4. Systemic threat analysis and information-sharing

a. **Identifying frequent perpetrators:** Identify users that are frequent perpetrators of threats to critical voices.

b. **Tracking online attacks:** Conduct regular analysis of the volume and reach of attacks, including understanding how online attacks spread quickly, using bots, graphics, memes, etc., as well as the way attacks may differentiate between those targeting women critical voices and those targeting men. The process should highlight risks specific to critical voices from groups in situations of vulnerability and marginalization.

c. **Behavioural nudges:** Use transparent nudges – real-time prompts – to discourage behaviour that harms human rights behaviour and prompt users to pause and reconsider potentially harmful language before posting. Nudges can be applied holistically across posts, comments, replies and direct messages.

d. **Quarantine systems for grey-area content**: Platforms should introduce quarantine systems to temporarily isolate content that falls into grey areas – allowing for thoughtful moderation before public exposure. This system would allow for careful review and contextual evaluation before a final decision is made. By quarantining content, this approach reduces immediate potential harm while avoiding over-censorship of content and thus preserving safety and freedom of expression.

e. **Deterrent public messaging:** Providers should consider sharing their risk assessment playbooks, or elements thereof, to deter perpetrators with a strong message that companies are ready and willing to act against threats to human rights on their services.

5. Training and resources for critical voices

a. **Digital security training**: Provide resources and tailored training on digital security good practices, including how to protect personal information and secure communication channels. Training can be conducted by non-governmental organizations (NGOs) and media organizations, journalism associations or unions, as appropriate. Schools of journalism could also provide training activities or curricula dedicated to digital security for aspiring journalists or established ones willing to gain more knowledge about the issue.[38]

b. **Media and information literacy (MIL), and digital education**: Provide training on digital literacy for critical voices to enhance their understanding of online platforms, algorithms, and the broader digital ecosystem. Such education empowers critical voices to navigate the digital landscape and engage with digital tools more critically and effectively, while promoting ethical practices in the digital space.

c. **Crisis response kits**: Offer digital tools and crisis response kits that guide critical voices under attack, such as secure ways to communicate and tools to counteract doxxing.

6. Training for the staff of digital platforms

Providing training for the staff of digital platforms, including their contractors, to enhance their understanding of the risks critical voices may face when using their products and services. These training sessions should aim to enhance their understanding of their human rights responsibilities, while raising awareness of the unique threats critical voices encounter and equipping staff with the skills to:

a. Recognize risk factors when designing, promoting, using, and deploying digital products and services.

b. Identify, assess, and mitigate human rights risks that could affect critical voices.

c. Respond effectively to complaints and take proactive measures to address concerns.

This approach ensures staff are well-informed and capable of supporting the safety and rights of critical voices in the digital space.

Similar to training programmes for critical voices, these sessions could be conducted in collaboration with CSOs, international organizations and other expert bodies.

Enhancing human rights defenders and journalists' capacities: integrating capacity building into risk management

Digital platforms should also work to strengthen the capacities of human rights defenders and journalists to effectively manage the risks they face. Enhancing these capacities can help reducing overall risk. Thus, some mitigation measures should be geared towards building these capacities. For instance, in the context of online attacks, mitigation strategies should include equipping journalists with digital security tools and resources. This empowers them to safeguard their online presence and handle threats more effectively. Additionally, platforms should develop and enforce community guidelines specifically addressing harassment, ensuring that these guidelines are well-integrated and actionable. A 2023 OHCHR pilot study[39] found that: 'Interviewees

38 See for example, UNESCO. Foley, M., Arthurs, C., Abu-Fadil, M. 2017. *Model course on safety of journalists: a guide for journalism teachers in the Arab States*. Lesson 6. https://unesdoc.unesco.org/ark:/48223/pf0000248297.
See also UNESCO, Jaakola, M. 2023. *Reporting on artificial intelligence: A handbook for journalism educators*. https://unesdoc.unesco.org/ark:/48223/pf0000384551.

39 OHCHR. 2023. Results overview: Pilot study on experiences with social media and communication platforms in MENA and East Africa regions. https://www.ohchr.org/sites/default/files/documents/issues/civicspace/Results-overview-of-pilot-study-on-experiences-with-social-media-and-communication-platforms-in-MENA-and-East-Africa-regions-June-2023.pdf.

emphasized the need for companies to provide direct guidance and support for human rights defenders and journalists using their platforms, including through training materials and opportunities for groups operating in different contexts to improve their risk awareness and cyber protection skills.' The UNESCO Guidelines touch upon this aspect under Principle 4, titled "Platforms make information and tools available for users." Although this principle does not explicitly reference human rights defenders and journalists, it underscores the importance of providing accessible information and tools for all users. Platforms are encouraged to offer their terms of service in the primary languages of their operating regions and to ensure that all users, including groups in situation of heightened risk or situation of marginalization, such as children and persons with disabilities, can access and understand this information. By doing so, platforms contribute to enhancing the overall capacities of their users, including journalists, to navigate and respond to online risks effectively.

Moreover, in alignment with the UN Guiding Principles and UNESCO Guidelines, digital platforms should emphasize the importance of MIL as a crucial component of their risk management approach. MIL equips users with the skills needed to critically assess and navigate the online environment. The Guidelines advocate for platforms to embed MIL into their product development processes, ensuring that teams are trained on online safety and critical media evaluation from a user empowerment perspective. Platforms should implement robust internal and external monitoring mechanisms to evaluate the effectiveness of these training programmes and to continuously refine their approach based on user feedback and emerging risks. This approach not only enhances human rights defenders and journalists' ability to navigate online threats but also fosters a more informed, diverse and resilient online information space.

7. Collaboration with international organizations and governments

a. **Collaborate with international organizations** and mechanisms, such as UNESCO, OHCHR and the UN Special Rapporteur on Freedom of Expression and Opinion, and regional mechanisms, such as the Inter-American Commission Special Rapporteur on Freedom of Expression, the African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information in Africa, the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, and the Council of Europe when developing preventive measures. Engaging these entities can provide valuable expertise in human rights, ensuring that the measures are aligned with international standards and current good practices. Their input can help to strengthen the guidance's effectiveness in addressing risks to journalists and promoting a safer digital environment.

b. **Engagement with government authorities:** Engagement with government authorities may in some cases be necessary to ensure access to justice and accountability, but such engagement may also represent a risk factor itself that can lead to violation of the rights of critical voices. In many situations, collaboration between digital platforms and governments have resulted in serious risks to the protection of critical voices and the right to freedom of expression. Critically assessing the context for such engagement and the conditions on which such engagement takes place is therefore indispensable.

As previously highlighted, the responsibility to respect human rights exists 'over and above compliance with national laws and regulations protecting human rights.' Thus, while businesses are expected to comply with local laws, they are also expected to respect internationally recognized human rights, which may entail operating to a higher standard in such contexts.

Any engagement with government authorities, including in response to government requests, should be guided by the context assessment outlined above and, where applicable, should take into account the consent of victims. A central element of this is the assessment of the degree of rule of law in a given context and at a given time, and the type of risks associated with weak levels of the rule of law.

Such scenarios can arise in the context of investigations and judicial proceedings. Any cooperation with authorities for the provision of data or content, should be based on law and be subject to prior judicial authorization, and notification to the user. While recognizing that collaboration may be important to identify and sanction perpetrators and can involve

establishing clear communication channels to report and swiftly address threats such as harassment, doxxing, or cyberattacks, any state-platform cooperation must be aligned with rule of law safeguards and international human rights standards. In addition, in cases where authorities' requests do not conform to these standards, service providers could challenge such requests through engagement or, when necessary, by seeking judicial review.

8. Redress systems

a. **Establish simple, easily accessible and user-friendly reporting mechanisms** to allow people to easily report harassment. Commit to developing a rapid response and early warning mechanism of any user reports for example by creating automatic documentation features that are built into their site.

b. **Strengthen methods of reporting**, including more robust reporting mechanisms that identify falsified content. Building content reporting methods, including crowd-based ones that encourage reporting and identifying deepfake content will reduce the number of vectors that could be used to propagate TFGBV with the use of generative AI. Reporting mechanisms should be accessible to all and consider procedures to guard against their misuse[40] in bad faith. In addition, such mechanisms should be designed to protect groups and voices at risk, responding to regional and linguistic diversity.

c. **Reporting mechanisms**

> Even though mitigation measures and capacity enhancement strategies are designed to reduce risks that could harm critical voices' safety online, no risk guidance can guarantee that adverse impacts on rights will not occur. Therefore, digital platforms should develop robust reporting mechanisms that allow online users, including critical voices, to formally report policy violations as well as other risks that may not be covered by existing policies, such as threats, harassment, violence, or infringements of their human rights.
>
> These mechanisms are central to ensuring the safety and security of critical voices in both online and offline environments. According to the UN Guiding Principles and UNESCO Guidelines, digital platforms should design these mechanisms to provide online users with clear and accessible channels for communication or reporting. These responses should be rapid and effective, and digital platforms should ensure that appropriate actions are taken to mitigate risks and support critical voices in their vital role of providing information to the public.

d. **Respond swiftly to reports of direct threats and attacks**[41] and analyse the accounts generating the text, images, video, among other types of content, as well as network accounts disseminating or also engaging with those. Increasing the speed of response and implementing features that will reduce the number of users that engage with the content can help in greatly limiting the harm done.

e. **Streamline reporting mechanisms and communication channels.** Digital platforms should implement clear, efficient, and accessible channels for reporting issues, with a gender-sensitive approach that addresses the unique challenges faced by women critical voices and which is geared towards nuances in specific regions and languages. These mechanisms should prioritize inclusivity by providing multilingual support, anonymous reporting options, and features that accommodate survivors of gender-based harassment or violence.

To ensure timely responses, platforms should designate specific points of contact for each country or region, trained to handle gender-sensitive cases with empathy and professionalism. This approach allows those that have been targeted, particularly women and marginalized groups, to connect with responsible representatives directly and safely when problems arise. By optimizing these mechanisms with a gender lens, platforms can foster a more effective and responsive system for crisis management, ensuring swift action, accountability, and protection.

40 Users should have the possibility to appeal against content removals through a counter-notification procedure.

41 See OSCE. Posetti, J., Maynard, D., Shabbir, N. 2023. *Guidelines for Monitoring Online Violence Against Female Journalists.* https://www.osce.org/files/f/documents/b/0/554098_1.pdf. The report provides 15 key indicators for online violence (against female journalists) escalation.

Criteria for effective grievance mechanisms:

Legitimate: enabling trust from the stakeholder groups for whose use they are intended and being accountable for the fair conduct of grievance processes.

Accessible: being known to all stakeholder groups for whose use they are intended and providing adequate assistance for those who may face barriers to access.

Predictable: providing a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring implementation.

Equitable: seeking to ensure that aggrieved parties have reasonable access to sources of information, advice and expertise necessary to engage in a grievance process on fair, informed and respectful terms.

Transparent: keeping parties to a grievance informed about its progress and providing sufficient information about the mechanism's performance to build confidence in its effectiveness and meet any public interest at stake.

Rights-compatible: ensuring that outcomes and remedies meet with internationally recognized human rights.

A source of continuous learning: drawing on relevant measures to identify lessons for improving the mechanism and preventing future grievances and harms.

Operational-level mechanisms should also be **based on engagement and dialogue**: consulting the stakeholder groups for whose use they are intended on their design and performance and focusing on dialogue as the means to address and resolve grievances.

9. Secure collaboration tools

Protected collaboration platforms: Ensure that platforms used by media, civil society and other relevant organizations for collaboration are fortified against unauthorized access and breaches.

10. Other

Create solutions for identifying falsified content and deepfakes: Leveraging creative solutions to identify 'fake' content, such as automatically checking for watermarks and labelling images can help reduce the number of attacks through generative AI. Such measures have limitations, including limited resilience to manipulation and risk of false positives, and may also not be feasible for smaller actors with limited technical capacities.

Considering the nature of online threats and the ever-evolving digital landscape, digital platforms should consider the following key areas to strengthen their human rights risk management strategies:

OBJECTIVES	ACTIONS FROM THE DIGITAL PLATFORMS
REGULAR REVIEW AND EVALUATION	Regularly review and evaluate the effectiveness of human rights impact assessment processes and mitigation measures, especially in extreme cases such as incidents resulting in deaths or severe harassment preventing them from exercising their duties.
ESTABLISHMENT OF A MONITORING DATABASE	Establish a database or utilize the existing ones to track and analyse threats, technological attacks, or violence against critical voices, ensuring the inclusion of gender-disaggregated data. This initiative should also encourage cross-platform cooperation to better understand and mitigate risks such as pile-on online violence and coordinated campaigns, fostering a more comprehensive and effective response to these threats.

DATA SHARING AND TRANSPARENCY WITH INTERNATIONAL ORGANIZATIONS AND CIVIL SOCIETY ORGANIZATIONS	Share anonymised data about risks with international organizations, such as UNESCO, OHCHR and civil society organizations, as well as independent oversight bodies, for instance, during electoral processes (in line with greater transparency from digital platforms).[42]
TRANSPARENCY FOR EFFECTIVE RISK MANAGEMENT[43]	Enhance transparency regarding platform operations and decision-making. Clearly communicate how data on threats and attacks against critical voices is collected, analysed, and used to inform safety measures. Make information on platform responses to reported incidents accessible to users and stakeholders. Ensure critical voices are aware of tools and practices available to protect themselves, fostering trust and enabling better risk management.
TRANSPARENCY OF THE RISK ASSESSMENT APPROACH	If the risk analysis indicates that sharing details about the assessment could further endanger the safety of critical voices, platforms should consider withholding or limiting such disclosures to protect those at risk. However, such measures should be considered exceptional and not the standard approach, ensuring transparency is maintained as much as possible while upholding safety and human rights standards.

RESPONSIVE MEASURES

The following sections outline examples of mitigation measures that digital platforms could implement to **reduce the severity of impact**. These measures aim to be practical, adhere to good practices and human rights standards, and ultimately enhance the safety and security of critical voices online.

1. Content moderation and policy enforcement

- **Robust content moderation:** Enhance content moderation by scaling both human and AI-based oversight during high-risk periods. Employ additional moderators to maintain a balanced workload, provide targeted training to strengthen their expertise, and ensure moderators are native speakers with the necessary cultural and contextual understanding. These steps enable swift and effective review of flagged content.
- **Anti-harassment policies:** Strengthen and consistently enforce policies against harassment, and threats. Ensure these policies are well-publicized and consistently applied.
- **Doxxing prevention:** Implement strict policies against doxxing. Automatically blur sensitive information like addresses or phone numbers shared publicly.
- **Swift account action:** Quickly suspend or ban accounts repeatedly and frequently used to attack critical voices, particularly in coordinated attacks, always bearing in mind the principles of legality, legitimacy and necessity/proportionality.
- **Fact-checking:** Implement systems for fact-checking. Prioritize fact-checking on the subject and consider appropriate measures to reduce the virality of such messages until the provenance is established.
- **Implement crowd-sourced fact-checking**: Leverage collective intelligence to verify claims quickly and accurately. This involves developing a structured platform where users can flag, review, and rate information, with expert oversight

42 See UNESCO Guidelines: Principle 3. 'Platforms are transparent.'
43 In line with Principle 3 of the Guidelines, specifically 'Meaningful transparency' and 'Data access for research purposes.'

to ensure credibility.

2. Support and protection

- **Dedicated reporting user-friendly channels:** Create a specific email address or designate a focal point where journalists, human rights defenders and other critical voices can report online harassment. Ensure the reviewer has the necessary security clearance and knowledge about the violence critical voices, particularly women, face in order to take appropriate action. Implement strong privacy and data protection measures to safeguard the personal information of critical voices using these channels, ensuring that reports are handled confidentially and securely, with data access strictly limited to authorized personnel.
- **Appeal mechanisms:** Ensure that critical voices can swiftly appeal if they believe their accounts have been unjustifiably suspended or labelled by platforms and that these appeals are swiftly reviewed.
- **Metadata protection:** Systematically deny requests for critical voices' metadata information from governments or non-state actors in contexts with a high level of impunity or violent threats against critical voices. Exceptions should only be made if the requests strictly adhere to a thorough and transparent due process of law, with explicit considerations to safeguard journalistic sources and protect against any potential harm to critical voices.
- **Technical vulnerabilities:** Inform users about technical vulnerabilities that could risk them becoming victims of unlawful surveillance tools, such as 0-click spyware. Take measures and policies to detect user exposure to unlawful surveillance.
- **Preserve and securely safeguard potentially critical content** for law enforcement and research purposes.
- **Collaborate with international organizations and civil society organizations (CSOs):** These collaborations can facilitate coordinated efforts to provide timely support, share expertise, and implement targeted interventions when risks materialize.

3. Training and capacity-building

- **Moderator and safety team training:** Provide comprehensive training not only for content moderation teams but also for all teams involved in safety, protection, risk assessment, and media-related issues. Training should cover MIL, local contexts, languages, and cultural sensitivities. Ensure diverse representation within these teams and establish clear processes for identifying, addressing, and correcting biased practices.

4. Transparency and accountability

- **Transparency in content removal:** Publish information about removed content based on laws that require platforms to give information to authorities, unless a specific gag order prevents this. Similarly, government requests to remove content should be disclosed.
- **Security audits and independent oversight:** Conduct regular security audits and vulnerability assessments to identify and address potential weaknesses. Develop anti-phishing tools to detect and block suspicious messages and links. Conduct independent external audits to improve oversight and increase confidence with users and regulators.
- **Secure storage:** Use secure storage solutions with encryption to protect sensitive information, including critical voices' personal and professional data.

5. Remedies

Critical voices whose human rights are violated by digital platforms' services and products should be afforded remedies.

The right to an effective remedy for human rights violations is enshrined in international human rights law.[44]

Digital platforms should establish comprehensive remediation mechanisms to address grievances and ensure accountability in their operations. These mechanisms can take various forms to provide effective remedies to affected parties:

- Involving institutions linked to government entities that can offer remedies through regulatory oversight or dispute resolution. Examples include regulators, ombudspersons, inspectorates, public complaints bodies, National Contact Points under the OECD *Guidelines for Multinational Enterprises on Responsible Business Conduct*.[45]
- Offering additional alternative pathways for remediation. Alternative pathways for remediation are typically developed and managed by private actors, such as individual companies, industry associations, or multi-stakeholder groups. In addition to creating new mechanisms, support – whether financial or otherwise – should also be extended to strengthen existing pathways and foster cooperation among these actors. Governance mechanisms should determine the appropriate type of remedy to apply, such as restitution, compensation, rehabilitation, satisfaction, or guarantees of non-repetition. Encourage the development of bystander, peer, and organizational support systems, and promote collaboration between institutions, multi-stakeholder groups, and bystander networks to establish multi-layered safety mechanisms.

4. Review and update

According to the UN Guiding Principles,[46] the UNESCO Guidelines, and B-Tech foundational papers, digital platforms should – as part of their ongoing due diligence – regularly review and update their assessments of human rights risk, particularly in response to significant changes in operations, services, or the design and development of new products. Also, in the cases of elections, crises, emergencies and conflict.

Companies need to evaluate the effectiveness of their actions to address human rights risks. The Guiding Principles set the expectation that 'In order to verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their response. Tracking should: (a) be based on appropriate qualitative and quantitative indicators; and (b) draw on feedback from both internal and external sources, including affected stakeholders.'[47]

Digital platforms are expected to report their assessment findings through appropriate governance channels and establish a review cycle for risk assessments. It is important for platforms to recognize key moments and triggers that may necessitate a new risk assessment.

Reporting on risk is a key element of effective risk management. Accurate and timely reporting through appropriate governance channels enhances organizational oversight and leads to improved risk management outcomes.

Monitoring the effectiveness of implemented measures is crucial for ongoing risk management, thus, to reduce the severity of harm to critical voices. Maintaining an up-to-date assessment is a necessary step. When reviewing and updating an existing assessment, digital platforms should incorporate new evidence to refine the most recent assessment.

In addition to digital platforms, it is essential that independent stakeholders, such as civil society organizations, media organizations, and media regulatory authorities, also participate in reviewing the human rights risk assessments implemented by companies. As underlined by the International Press

---

44 See in particular, article 8 of the Universal Declaration of Human Rights. 1948. https://www.ohchr.org/en/human-rights/universal-declaration/translations/english; and article 2 of the International Covenant on Civil and Political Rights. 1966. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.
See further United Nations General Assembly Basic Principles and *Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law*. 2005. https://www.ohchr.org/en/instruments-mechanisms/instruments/basic-principles-and-guidelines-right-remedy-and-reparation.

45 OECD. 2023. *Guidelines for Multinational Enterprises on Responsible Business Conduct*. 2023. https://www.oecd.org/en/publications/oecd-guidelines-for-multinational-enterprises-on-responsible-business-conduct_81f92357-en.html.

46 See article 17 'Human rights due diligence.'

47 See the OHCHR. 2023. B-Tech Foundational Paper 'Key Characteristics of Business Respect for Human Rights.' https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.

Institute (IPI) in its 'Protocol for newsrooms to support journalists targeted with online harassment,'[48] newsrooms should not only keep an eye on reported cases of online harassment, but they should also reassess the safety and support mechanisms to protect journalists from online harassment.

This external oversight ensures that the assessments are not only comprehensive and aligned with current good practices and international human rights standards but also holds digital platforms accountable for their obligations. Engaging diverse stakeholders helps mitigate the risk of negligence or superficial assessments, preventing digital companies from avoiding their due diligence responsibilities. This collaborative approach strengthens the overall integrity of the human rights impact assessment process and helps combat impunity, particularly in cases where harm to critical voices is involved.

CONCLUSION

The evolving digital landscape presents both opportunities and challenges for the safety of critical voices worldwide. As the threats facing human rights defenders and journalists become increasingly shaped by the digital world, it is imperative that digital platforms recognize their pivotal role in both contributing to and mitigating these risks.

Digital platforms are not merely passive environments but active participants in the risk equation. They are central to addressing the safety challenges faced by critical voices such as human rights defenders and journalists by implementing effective mitigation measures and enhancing their capacities. Adhering to international human rights law and standards of responsible business conduct according to the UN Guiding Principles and leveraging the UNESCO Guidelines, this HRIA Guidance offers a path for digital platforms to safeguard critical voices.

To effectively address these challenges, this Guidance outlines several key strategies. In this regard, digital platforms should implement robust HRIA processes.

Enhancing the capacities of critical voices is another crucial aspect addressed in this Guidance. Platforms should provide tools and resources to support critical voices in navigating digital threats.

Furthermore, this Guidance stresses the importance of establishing solid and efficient reporting mechanisms. Digital platforms should create clear and accessible channels for critical voices to report threats and harassment, ensuring that responses are prompt and effective.

However, protecting critical voices is not solely the responsibility of digital platforms, and should be considered in the context of the broader obligations of States to respect, protect and promote human rights. This Guidance also emphasizes the need for a multi-stakeholder approach to enhance their safety. Collaboration among digital platforms, journalists, media organizations, civil society organizations, international organizations and governments is crucial to creating an enabling environment for critical voices to do their work.

UNESCO and OHCHR are committed to overseeing the implementation of human rights due diligence processes for the protection of critical voices online and fostering a multistakeholder dialogue that enables companies to analyse context, assess human rights risks, and implement effective mitigation measures.

48 International Press Institute (IPI). 2020. *Protocol for newsrooms to support journalists targeted with online harassment.* https://newsrooms-ontheline.ipi.media/wp-content/uploads/2020/02/IPI_newsrooms_protocol_address_online_harassment_ok_022020.pdf.

RESOURCES

1. International organizations

UNESCO

- *2024. Press and Planet in Danger: Safety of Environmental Journalists; Trends, Challenges and Recommendations.* https://unesdoc.unesco.org/ark:/48223/pf0000389501.
- *Cherian, G. 2024. Covering Hate Speech: A Guide for Journalists.* https://unesdoc.unesco.org/ark:/48223/pf0000392378.
- *2023. Guidelines for the Governance of Digital Platforms: safeguarding freedom of expression and access to information through a multi-stakeholder approach.* https://unesdoc.unesco.org/ark:/48223/pf0000387339.
- *Berger, G., Gillwald, A., Orembo, E., Diouf, D. Garcia, J.M. 2023. Platform Problems and Regulatory Solutions: Findings from a Comprehensive Review of Existing Studies and Investigations.* https://unesdoc.unesco.org/ark:/48223/pf0000385813.
- *2023. Data Sharing to Foster Information as a Public Good: The Case of Media Viability and Safety of Journalists in the Digital Ecosystem.* https://unesdoc.unesco.org/ark:/48223/pf0000387896.
- *Chowdhury, R., Lakshmi, D. 2023. «Your Opinion Doesn't Matter, Anyway»: Exposing Technology-Facilitated Gender-Based Violence in an Era of Generative AI.* https://unesdoc.unesco.org/ark:/48223/pf0000387483.
- Jaakkola, M. 2023. *Reporting on Artificial Intelligence: A Handbook for Journalism Educators.* https://unesdoc.unesco.org/ark:/48223/pf0000384551.
- *Posetti, J., Bontcheva, K. 2022. The Chilling: Recommendations for Action Responding to Online Violence Against Women Journalists.* https://unesdoc.unesco.org/ark:/48223/pf0000383788.
- *Posetti, J., Shabbir, N., Maynard, D., Bontcheva, K. Aboulez, N. 2021. The Chilling: Global Trends in Online Violence Against Women Journalists – Research Discussion Paper.* https://unesdoc.unesco.org/ark:/48223/pf0000377223.
- *2021. Windhoek +30 Declaration: Information as a Public Good – World Press Freedom Day.* https://unesdoc.unesco.org/ark:/48223/pf0000378158.
- *Puddephatt, A. 2021. Letting the Sun Shine In: Transparency and Accountability in the Digital Age.* https://unesdoc.unesco.org/ark:/48223/pf0000377231.
- *Posetti, J., Aboulez, N., Bontcheva, K., Harrison, J., and Waisbord, S. 2020. Online Violence Against Women Journalists: A Global Snapshot of Incidence and Impacts.* https://unesdoc.unesco.org/ark:/48223/pf0000375136.
- *Smyth, F. 2020. Safety of Journalists Covering Protests: Preserving Freedom of the Press During Times of Turmoil.* https://unesdoc.unesco.org/ark:/48223/pf0000374206.
- Foley, M., Arthurs, C., Abu-Fadil, M. 2017. *Model Course on Safety of Journalists: A Guide for Journalism Teachers in the Arab States.* https://unesdoc.unesco.org/ark:/48223/pf0000248297.
- *2017. Executive Board Decision on the Safety of Journalists (201 EX/Decision 5.I.I).* https://unesdoc.unesco.org/ark:/48223/pf0000248900.
- *2017. Executive Board Decision on the Safety of Journalists (202 EX/Decision I.K).* https://unesdoc.unesco.org/ark:/48223/pf0000259824.

OHCHR

- 2025. *Online Platform Governance & Human Rights*. https://www.ohchr.org/en/documents/tools-and-resources/online-platform-governance-human-rights.
- 2023. B-Tech Foundational Paper 'Key Characteristics of Business Respect for Human Rights.' https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf.
- *2024. Civic Space and Tech. Brief on Encryption.* https://www.ohchr.org/en/documents/tools-and-resources/civic-space-tech-brief-encryption-0.
- 2024. Civic Space and Tech. Brief on Hacking and Spyware. https://www.ohchr.org/en/documents/tools-and-resources/civic-space-tech-brief-hacking-and-spyware.
- *2023. Pilot Study on Experiences with Social Media and Communication Platforms in MENA and East Africa Regions.* https://www.ohchr.org/sites/default/files/documents/issues/civicspace/Results-overview-of-pilot-study-on-experiences-with-social-media-and-communication-platforms-in-MENA-and-East-Africa-regions-June-2023.pdf.
- *2021. B-Tech Foundational Paper: Access to Remedy and the Technology Sector – Basic Concepts and Principles.* https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
- *2021. B-Tech Foundational Paper: Access to Remedy and the Technology Sector – A "Remedy Ecosystem" Approach.* https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-ecosystem-approach.pdf.
- *2021. B-Tech Foundational Paper: Designing and Implementing Effective Company-Based Grievance Mechanisms.* https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-company-based-grievance-mechanisms.pdf.
- *2021. B-Tech Foundational Paper: Access to Remedy and the Technology Sector – Understanding the Perspectives and Needs of Affected People and Groups.* https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/access-to-remedy-perspectives-needs-affected-people.pdf.
- *2020. B-Tech Foundational Paper: Identifying and Assessing Human Rights Risks Related to End-Use.* https://www.ohchr.org/sites/default/files/Documents/Issues/Business/B-Tech/identifying-human-rights-risks.pdf.
- 2012. United Nations Plan of Action on the Safety of Journalists and the Issue of Impunity. 2012. https://www.ohchr.org/sites/default/files/documents/issues/journalists/2023-01-31/un-plan-on-safety-journalists_en.pdf.
- 2011. United Nations Guiding Principles on Business and Human Rights. www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

UNITED NATIONS

- *2024. Global Digital Compact.* https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf.
- *2022. The Right to Privacy in the Digital Age. (A/HRC/51/17).* https://documents.un.org/doc/undoc/gen/g22/442/29/pdf/g2244229.pdf.

UN WOMEN

- 2024. UN Women. *Cybersecurity Threats, Vulnerabilities and Resilience Among Women Human Rights Defenders and Civil Society in South-East Asia.*https://unu.edu/sites/default/files/2024-05/Cybersecurity%20Threats%2C%20Vulnerabilities%20and%20Resilience%20Among%20Women%20Human%20Rights%20Defenders%20and%20Civil%20Society%20in%20South-East%20Asia.pdf.

OECD

- 2023. *Guidelines for Multinational Enterprises on Responsible Business Conduct.* https://www.oecd.org/en/publications/oecd-guidelines-for-multinational-enterprises-on-responsible-business-conduct_81f92357-en.html.

OSCE

- Posetti, J, Maynard, D. Shabbir, N. 2023. Guidelines for Monitoring Online Violence Against Female Journalists. https://www.osce.org/files/f/documents/b/0/554098_1.pdf.

2. European Union

- 2022. *Digital Services Act.* Article 34(1) 'Risk Assessment.' https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.
- 2022. Digital Services Act. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065.
- Transparency reports of the Digital Services Act. https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency.
- 2020. Council of Europe. *Recommendation CM/Rec (2020) of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems* .https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016809e1154.

3. Civil society organizations

- Search for Common Ground, Integrity Institute, and Council on Technology & Social Cohesion. 2025. *Prevention by Design: A Roadmap for Tackling Technology-Facilitated Gender-Based Violence at the Source.* https://techandsocialcohesion.org/wp-content/uploads/2025/03/Prevention-by-Design-A-Roadmap-for-Tackling-TFGBV-at-the-Source.pdf.
- PEN America. 2024. *The Power of Peer Support: Helping Journalists Persevere in the Face of Online Abuse.* https://pen.org/report/peer-support/.
- UltraViolet. 2021. New Report *Card Grades Social Media Platforms on Handling of Harassment, Hate Speech, Misogyny, Disinformation.* https://weareultraviolet.org/wp-content/uploads/2021/11/Social-media-fails-women.pdf.
- International Press Institute (IPI). 2020. *Protocol for Newsrooms to Support Journalists Targeted with Online Harassment.* https://newsrooms-ontheline.ipi.media/wp-content/uploads/2020/02/IPI_newsrooms_protocol_address_online_harassment_ok_022020.pdf.
- 2020. The Danish Institute for Human Rights. *Guidance on Human Rights Impact Assessment of Digital Activities.* www.humanrights.dk/files/media/document/A%20HRIA%20of%20Digital%20Activities%20-%20Introduction_ENG_accessible.pdf.
- #ShePersisted. https://she-persisted.org/our-work/research-and-thought-leadership/.
- Frontline Defenders. 2016. *Workbook on Security: Practical Steps for Human Rights Defenders at Risk.* https://www.frontlinedefenders.org/en/resource-publication/workbook-security-practical-steps-human-rights-defenders-risk.

4. International conventions

- 1976. OHCHR. *International Covenant on Civil and Political Rights.*www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

5. Other sources

- 2023. Ofcom. *Protecting People from Illegal Harms Online – Annex 5: Service Risk Assessment Guidance.* www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/annex-5-draft-service-risk-assessment-guidance/?v=330403.
- 2023. Online Safety Act. http://www.legislation.gov.uk/ukpga/2023/50/2024-08-23.
- A Tech Accord to Combat Deceptive Use of AI in 2024 Elections. 2024.

ABOUT THE CONTRIBUTORS

The HRIA Guidance is a working tool designed to support the implementation of the UNESCO *Guidelines for the Governance of Digital Platforms* and the *United Nations Guiding Principles on Business and Human Rights* of the OHCHR B-Tech Project.

The Guidance is the product of collaborative efforts led by OHCHR and UNESCO, specifically involving:

OHCHR

Staff members of the Rule of Law, Governance and Civic Space Branch of the Thematic Engagement Division.

UNESCO

- Sylvie Coudray, Director for Freedom of Expression, Media development and Media and Information Literacy (CI/FMD) and Secretary of the International Programme for the Development of Communication (IPDC)
- Andrea Cairola, Chief of Section, Freedom of Expression and Safety of Journalists, a.i.
- Ana Cristina Ruelas, Senior Programme Specialist
- Ophélie Kukansami Léger, Associate Project Officer and
- Daria Kovaleva, Consultant

The first draft of the HRIA Guidance was developed by Andrew Puddephatt, following the thematic consultation held in Copenhagen, Denmark, in 2022, organized as part of the 10th anniversary of the United Nations Plan of Actions on the Safety of Journalists and the Issue of Impunity.

Participants of the various consultations also provided crucial insights and contributions, namely attendees of the following events:

- Hybrid Thematic Consultation on the Safety of Journalists in a Digital Age, 13 September 2022, Copenhagen, Denmark.
- Consultation held within the framework of the Forum on Internet Freedom in Africa (FIFAfrica), 25–27 September 2024, Dakar, Senegal.
- Latin American Conference on Investigative Journalism (COLPIN), 23–26 October 2024, Madrid, Spain.
- Consultation held in the context of the 10th International Day to End Impunity for Crimes against Journalists, 6–7 November 2024, Addis Ababa, Ethiopia.
- World Press Freedom Day session, 6 May 2025, Brussels, Belgium.

The HRIA Guidance has been shaped and informed by the insights of numerous external reviewers from all geographic regions, including civil society organizations, digital platforms, journalists and media professionals.

Funding for this work was provided by the Multi-Donor Programme on Freedom of Expression and the Safety of Journalists.

PROTECTING CRITICAL VOICES

Guidance for Human Rights Impact Assessment on Digital Platforms

A wide range of digital technologies operate within platforms' ecosystems — from search and discovery services to advertising systems, messaging tools, and generative artificial intelligence (AI) content production. Risks to critical voices can emerge from multiple sources, including the design of these tools or services, as well as from the platforms' own policies and practices.

This Guidance, developed jointly by UNESCO and the Office of the United Nations High Commissioner for Human Rights (OHCHR), seeks to help companies identify, assess, and address human rights risks associated with digital platforms. Emphasizing collaboration among stakeholders, it promotes responsible digital governance to safeguard freedom of expression and protect journalists and human right defenders across interconnected online and offline environments.

https://www.unesco.org/en/internet-trust?hub

internetconference@unesco.org

unesco

UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

9 789231 008375