Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms

INSIGHT REPORT

AUGUST 2023

Images: Getty Images, Unsplash

Contents

Disclaimer
This document is published by the World Economic Forum as a contribution to a project, insight area or interaction. The findings, interpretations and conclusions expressed herein are a result of a collaborative process facilitated and endorsed by the World Economic Forum but whose results do not necessarily represent the views of the World Economic Forum, nor the entirety of its Members, Partners or other stakeholders.

Foreword

Julie Inman Grant
Australian eSafety Commissioner

Adam Hildreth
Founder, Crisp, a Kroll business

Daniel Dobrygowski
Head, Governance and Trust, World Economic Forum

Minos Bantourakis
Head, Media, Entertainment and Sport Industry, World Economic Forum

The internet has transformed the world into a global village, connecting people from different corners of the world with ease and speed. However, it has also heightened various social harms, such as bullying and harassment, hate speech, disinformation and radicalization. The amplification of these harms has far-reaching consequences, affecting individuals, communities and societies.

While the internet is global in nature, harms can be highly local or context-specific: unique risks may arise in different countries or regions or in different communities. Factors such as cultural norms, legal systems and societal values influence how individuals perceive and respond to online threats. Within this context, it is important to acknowledge that digital safety requires a complex range of deliberations, balancing legal, policy, ethical, social and technological considerations. Digital safety decisions must be rooted in international human rights frameworks.[1]

To address the complex landscape of harms in online spaces, the World Economic Forum's Global Coalition for Digital Safety[2] is developing the Typology of Online Harms, as outlined in this report. The report intends to work towards creating a common terminology as well as a shared understanding when discussing online harms across jurisdictions. Moreover, it aims to facilitate conversations about online harms, but it does not set out to provide any severity ratings or to be used for regulatory compliance.

The Coalition, intending to foster cooperation on digital safety, has created this document through extensive research, collaboration and expert consultations. This typology complements other significant publications by the Coalition, including *Global Principles on Digital Safety*,[3] published in January 2023, which recognizes the need to advance a shared understanding of online harm issues. It also complements *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*,[4] launched in May 2023. Together, these publications offer valuable resources for policy-makers, industry leaders, civil society organizations, researchers and individuals, helping to address the issue of harmful content online comprehensively and in a rights-respecting manner.

Executive summary

The Typology of Online Harms aims to provide a foundational common language, facilitating multistakeholder and cross-jurisdictional discussions to advance digital safety.

Developed by a working group of the Global Coalition for Digital Safety, comprising representatives from industry, governments, civil society and academia, this typology serves as a foundation for facilitating multistakeholder discussions and cross-jurisdictional dialogues to find a common terminology and shared understanding of online safety.

The Typology of Online Harms is an integral part of the Toolkit for Digital Safety Design Interventions and Innovations, one of the key workstreams initiated by the Global Coalition for Digital Safety. This toolkit aims to define online harms and identify the potential technology, policy, processes and design interventions needed to advance digital safety in a rights-respecting manner. By aligning with the commitment to foster a shared understanding of online harm issues through a multistakeholder dialogue, as well as the call for governments to clearly distinguish between illegal content and content that is lawful but may be harmful as outlined in the *Global Principles on Digital Safety*,[5] the typology complements the Coalition's efforts to promote digital safety while respecting individuals' rights. It can also be effectively used in conjunction with *Digital Safety Risk Assessment In Action*,[6] as this typology provides a common language for categorizing and defining the various types of online harms that require assessment.

> **“ The typology takes into account both regional and local contexts. It recognizes the complex and interconnected nature of online safety, encompassing content, contact and conduct risks.**

Considering the global nature of the internet and the local implications of online harms, the typology takes into account both regional and local contexts. It recognizes the complex and interconnected nature of online safety, encompassing content, contact and conduct risks. While recognizing the value of contract as a fourth "C" to reflect risks for children in relation to commercialization and datafication,[7] this typology uses a "three C" framework to encompass online safety risks for a broad range of end users. It categorizes these harms, including threats to personal and community safety such as child exploitation and extremist content, harm to health caused by content promoting suicide or disordered eating and violations of dignity and privacy through bullying and harassment, doxxing and image-based abuse. Deception and manipulation, such as disinformation and manipulated media, are also addressed.

While this publication does not specifically cover emerging technologies such as the metaverse, Web3 or generative AI, it emphasizes the need for the ongoing development of processes that keep pace with technological advances and their societal impact. To this end, the Coalition will also look at developing a conceptual and comprehensive framework to ensure that the approach to online harms is future-proof.

In addition to the Typology of Online Harms, the Coalition is preparing two upcoming reports that will complement this work. The first report, *Risk Factors, Metrics and Measurement*, focuses on approaches to evaluate the risks of adverse impacts from online harms, as well as the benefits of mitigation actions, and the second, *Solution-Based Interventions*, draws on safety-by-design principles and best practices to provide a resource to assist companies in effectively identifying and reducing digital risks, preventing harm and promoting trust and safety.

1 Introduction

The Typology of Online Harms serves as a foundation to build a common terminology and shared understanding of the diverse range of risks that arise online, including in the production, distribution and consumption of content.

Online harms encompass various dimensions, including harm in **content** production and distribution, as well as harm in content consumption:

- Harm in the production of content – for example, where a person is physically harmed, and the abuse is recorded or streamed in order to create online material. This could include images or videos of murder, assault or the sexual abuse of adults or children.

- Harm in the distribution of content – for example, where an intimate image of a person is self-produced and shared voluntarily and is later shared and distributed online without their consent. That person may not have been harmed in the production of the content but is exposed to harm once their intimate image is shared. Similarly, victims who are the subject of abuse in the production of content can face compounded trauma when that content is distributed. Those who film, share or consume the content also risk being harmed. The person objectified in such content is also harmed because the distribution of that content can reinforce negative attitudes towards certain populations. Amplifying or resharing hateful comments about a minority group serves as an example that reinforces stereotypes towards the underrepresented group, perpetuating biases and inflicting further harm on these individuals.

- Harm in the consumption of content – for example, where a person is negatively affected as a result of viewing illegal, age-inappropriate, potentially dangerous or misleading content.

Online harms can also occur as a result of online interactions with others (**contact**) and through behaviour facilitated by technology (**conduct**).

The following typology of harms builds on existing and emerging international approaches to understanding and mitigating online harms, as listed in the Resources section of this document, and considers the need to address online harms in a rights-respecting way. Harms may be concurrent and intersecting, and their categorization is not always exclusive. For example, while image-based abuse is included under the heading of privacy, it can also relate to harms to personal safety and health and well-being. Similarly, child sexual exploitation and abuse is a threat to personal and community safety, harmful to health and well-being, a violation of dignity and an invasion of privacy. Online harms may also form part of a broader harm context that can occur across a range of technologies and include behaviours perpetrated online and offline, and risk vectors can overlap for certain harms. For example, while child sexual abuse material (CSAM) is listed as primarily a content risk, it may be produced and distributed as a result of contact or conduct, such as grooming and sexual extortion.

By framing online harms through a human rights lens, this typology emphasizes the impacts on individual users and aims to provide a broad categorization of harms to support global policy development. It notes that there are regional differences in how specific harms are defined in different jurisdictions and that there is no international consensus on how to define or categorize common types of harm. Considering the contextual nature of online harm, the typology does not aim to offer precise definitions that are universally applicable in all contexts.

In line with the Global Principles on Digital Safety, this typology underscores the importance of balancing different rights, acknowledging that all types of online harm have the potential to unlawfully deny individuals their right to participate and express themselves online.

At this stage it is largely up to individual online service providers to establish rules and guidelines for the types of activity and content that are or are not permitted on their platforms within community guidelines or terms of service. However, these can diverge significantly across services.

This typology can help companies, including those at the early stage, to understand the range of online harms that might occur to users of their services – as well as their impacts on victims, the different modes of abuse and factors that might contribute to harm. Moreover, the typology can aid governments by establishing a shared language to identify online harms and facilitate the efforts of civil society organizations seeking to participate in multistakeholder discussions that advocate for a safer online ecosystem. By using this typology, stakeholders from all sectors can promote a collaborative approach to address the challenges posed by online harms.

Harms to corporations and brands (e.g. copyright infringement) are not within the scope of this typology. The typology does not aim to prescribe actions to be taken in response to harms nor does it seek to assign severity ratings to harms. Furthermore, the typology is focused on online harms affecting individuals and society, but cannot be considered fully exhaustive in terms of all types of harms (e.g. animal cruelty). It does provide a common foundation for multistakeholder discussions to develop a shared terminology for and understanding of online harms. In determining appropriate interventions for content or conduct falling within any of these harm categories, the broader context of international human rights conventions needs to be considered, as do potential contextual exceptions for content and conduct that is newsworthy, educational, artistic or has other merits.

Possible harms that may affect future technology paradigms, including the metaverse and Web3, are not outlined in this report. However, as a next step, the Coalition will develop a future-proof framework on online harms to help different stakeholders keep pace with technological advances.

2 Typology of Online Harms

The typology recognizes the complex and interconnected nature of online safety, encompassing content, contact and conduct risks.

2.1 Threats to personal and community safety

Although online harms are commonly targeted at an individual, they have broader community and societal impacts. Similarly, societal harms have individual impacts and consequences.

The United Nations Convention on the Rights of the Child General Comment 25 asserts that: "State parties should take legislative and administrative measures to protect children from violence in the digital environment, including the regular review, updating and enforcement of robust legislative, regulatory and institutional frameworks that protect children from recognized and emerging risks of all forms of violence in the digital environment."[8]

Article 3 of the Universal Declaration of Human Rights states that: "Everyone has the right to life, liberty and security of person."[9]

a. Content risks

1. **Child sexual abuse material (CSAM)**

   Any representation by whatever means of a child engaged in real or simulated explicit sexual activities or any representation of the sexual parts of a child for primarily sexual purposes. While the laws of many countries continue to use the term "child pornography", there has been a global movement towards the use of the term "child sexual abuse material" (CSAM) to properly convey that sexualized material depicting or otherwise representing children is indeed a representation, and a form, of child sexual abuse.[10]

2. **Child sexual exploitation material (CSEM)**

   Content that sexualizes and is exploitative of the child, whether or not it shows the child's sexual abuse.

3. **Pro-terror material**

   Material that advocates engaging in a terrorist act because it counsels, promotes, encourages or urges engaging in a terrorist act, provides instruction on engaging in a terrorist act or directly praises engaging in a terrorist act in circumstances where there is a substantial risk that such praise might have the effect of leading a person to engage in a terrorist act.[11]

4. **Content that praises, promotes, glorifies or supports extremist organizations or individuals**

   Includes content that encourages participation in, or intends to recruit individuals to, violent extremist organizations – including terrorist organizations, organized hate groups, criminal organizations and other non-state armed groups that target civilians – with names, symbols, logos, flags, slogans, uniforms, gestures, salutes, illustrations, portraits, songs, music, lyrics or other objects meant to represent violent extremist organizations or individuals.

5. **Violent graphic content**

   Content that promotes, incites, provides instruction in or depicts acts including murder, attempted murder, torture, rape and kidnapping of another person using violence or the threat of violence. It is important to consider the various contexts in which this content may arise, including both condemning/informative purposes and in the context of documenting human rights abuses.

6. **Content that incites, promotes or facilitates violence**

   Includes content that contains direct and indirect threats of violence and intimidation.

7. **Content that promotes, incites or instructs in dangerous physical behaviour**

   Content that promotes, incites or provides instruction in activities conducted in a non-professional context that may lead to serious injury or death for the user or members of the public.

b. Contact risks

1. **Grooming for sexual abuse**

   When someone uses the internet to deliberately establish an emotional connection with a young person to lower their inhibitions, and make it easier to have sexual contact with them. It may involve an adult posing as a child in an internet application to befriend a child and encourage them to behave sexually online or to meet in person.

2. **Recruitment and radicalization**

   Includes posting or engaging with individuals with the purpose of recruiting individuals to a designated or dangerous organization.

c. Conduct risks

1. **Technology-facilitated abuse (TFA)**

   Using digital technology to enable, assist or amplify abuse or coercive control of a person or group of people.

2. **Technology-facilitated gender-based violence**

   A subset of technology-facilitated abuse that captures any act that is committed, assisted, aggravated or amplified by the use of information communication technologies or other digital tools, resulting in or likely to result in physical, sexual, psychological, social, political or economic harm or other infringements of rights and freedoms on the basis of gender characteristics.

d. Content/contact/conduct risks

1. **Child sexual exploitation and abuse (CSEA)**

   Can refer to content (e.g. CSAM), contact (e.g. grooming) and conduct (e.g. livestreaming).

2.2 Harm to health and well-being

Harmful online content and behaviour can be seriously damaging, especially for those most at risk. The social, emotional, psychological and even physical impacts of online harms can be immediate, experienced over time and/or enduring. They can also be experienced both online and offline.

Nevertheless, online platforms and tools not only harbour harmful content and behaviour but also provide a safe space for individuals to address these issues. This allows people to share experiences, raise awareness, access mental health resources and seek support from one another, both online and offline.

The United Nations Convention on the Rights of the Child General Comment 25 states that: "The use of digital devices should not be harmful, nor should it be a substitute for in-person interactions among children or between children and parents or caregivers. State parties should pay specific attention to the effects of technology in the earliest years of life, when brain plasticity is maximal and the social environment, in particular relationships with parents and caregivers, is crucial to shaping children's cognitive, emotional and social development. In the early years, precautions may be required, depending on the design, purpose and uses of technologies. Training and advice on the appropriate use of digital devices should be given to parents, caregivers, educators and other relevant actors, taking into account the research on the effects of digital technologies on children's development, especially during the critical neurological growth spurts of early childhood and adolescence."[12]

Article 12 (1) of the International Covenant on Economic, Social and Cultural Rights outlines the right to the "enjoyment of the highest attainable standard of physical and mental health".[13]

a. Content risks

1. **Material that promotes suicide, self-harm and disordered eating**

   Content that promotes suicidal or self-injurious behaviour. Includes content that promotes, encourages, coordinates or provides instructions on:

   – Suicide

   – Self-injury, including depictions of graphic self-injury imagery

   – Eating disorders, including expressing desire for an eating disorder, sharing tips or coaching on disordered eating, or encouraging participation in unhealthy body measurement challenges

2. **Developmentally inappropriate content**

   Includes children's access to pornography, particularly of a violent or extreme nature, and graphic, violent material.

2.3 Hate and discrimination

Online hate and discrimination can negatively affect a person's mental health, general well-being and online engagement. It can also, in the most extreme cases, lead to harassment and violence offline.

The United Nations Convention on the Rights of the Child General Comment 25 calls upon: "State parties to take proactive measures to prevent discrimination on the basis of sex, disability, socioeconomic background, ethnic or national origin, language or any other grounds and discrimination against minority and Indigenous children, asylum-seeking, refugee and migrant children, lesbian, gay, bisexual, transgender and intersex children, children who are victims and survivors of trafficking or sexual exploitation, children in alternative care, children deprived of liberty and children in other vulnerable situations."[14]

The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) recognize rights to equality and non-discrimination, which can include protections against violations of the right to safety. Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR) states that: "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[15]

a. Content risks

1. **Hate speech**

   Any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of their inherent/protected characteristics – in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, gender or other identity factor. Includes dehumanization, which targets individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease or any other non-human entity.

b. Conduct risks

1. **Algorithmic discrimination**

   A decision that results in the denial of financial and lending services, housing, insurance, education enrolment, criminal justice, employment opportunities, healthcare services or access to basic necessities, such as food and water.

   It is important to acknowledge that certain practices, such as age restrictions implemented by platforms, serve as protective measures to prevent harmful interactions between unrelated adults and teenagers. Therefore, in this as in other definitions, it is crucial to strike a balance, ensuring that while implementing such measures there is a commitment to upholding human rights principles, as emphasized in the Global Principles on Digital Safety.

2.4 Violation of dignity

The Universal Declaration of Human Rights Article 1 states that: "All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood."[16]

a. Conduct risks

1. **Online bullying and harassment**

   The use of technology to bully someone – to deliberately engage in hostile behaviour to hurt them socially, emotionally, psychologically or even physically. This can include abusive texts and emails; hurtful messages, images or videos; excluding others; spreading damaging gossip and chat; or creating fake accounts to trick or humiliate someone.

b. Contact risks

1. **Sexual extortion**

   Also called "sextortion", the blackmailing of a person with the help of self-generated images of that person in order to extort sexual favours, money or other benefits from them under the threat of sharing the material beyond the consent of the depicted person (e.g. posting images on social media). Often, the influence and manipulation typical of groomers over longer periods of time (sometimes several months) turns into a rapid escalation of threats, intimidation and coercion once the person has been persuaded to send the first sexual images of themself.[17]

2.5 Invasion of privacy

The United Nations Convention on the Rights of the Child General Comment 25 states that: "Privacy is vital to children's agency, dignity and safety and for the exercise of their rights. Children's personal data are processed to offer educational, health and other benefits to them. Threats to children's privacy may arise from data collection and processing by public institutions, businesses and other organizations, as well as from such criminal activities as identity theft. Threats may also arise from children's own activities and from the activities of family members, peers or others, for example, by parents sharing photographs online or a stranger sharing information about a child."[18]

a. Conduct risks

1. **Doxxing**

   The intentional online exposure of an individual's identity, personal details or sensitive information without their consent and with the intention of placing them at risk of harm.

2. **Image-based abuse**

   Sharing, or threatening to share, an intimate image or video without the consent of the person shown. An "intimate image/video" is one that, where there is a reasonable expectation of privacy, shows nudity, sexual poses, private activity such as showering or someone without the religious or cultural clothing they would normally wear in public.

2.6 Deception and manipulation

Article 25 of the International Covenant on Civil and Political Rights describes the right to free and fair elections. Article 12 of the International Covenant on Economic, Social and Cultural Rights outlines the right to health.

The United Nations Convention on the Rights of the Child General Comment 25 asserts that: "State parties should ensure that uses of automated processes of information filtering, profiling, marketing and decision-making do not supplant, manipulate or interfere with children's ability to form and express their opinions in the digital environment."[19]

a. Content risks

1. Disinformation and misinformation

Two distinct types of false or inaccurate information. Misinformation involves the dissemination of incorrect facts, where individuals may unknowingly share or believe false information without the intent to mislead. Disinformation involves the deliberate and intentional spread of false information with the aim of misleading others. Both can be used to manipulate public opinion, interfere with democratic processes such as elections or cause harm to individuals, particularly when it involves misleading health information. Includes gendered disinformation that specifically targets women political leaders, journalists and other public figures, employing deceptive or inaccurate information and images to perpetuate stereotypes and misogyny.

2. Deceptive synthetic media

Content that has been generated or manipulated via algorithmic processes (such as artificial intelligence or machine learning) to appear as though based on reality, when it is, in fact, artificial and seeks to harm a particular person or group of people. Includes deepfakes, which are extremely realistic – although fake – images, audio or video clips that show a real person doing or saying something that they did not actually do or say.

b. Conduct risks

1. Impersonation

Posing as an existing person, group or organization in a confusing or deceptive manner.

2. Scams

Dishonest schemes that seek to manipulate and take advantage of people to gain benefits such as money or access to personal details.

3. Phishing

The sending of fraudulent messages, pretending to be from organizations or people the receiver trusts, to try and steal details such as online banking logins, credit card details and passwords from the receiver.

4. Catfishing

The use of social media to create a false identity, usually to defraud or scam someone. People who catfish often make up fake backgrounds, jobs or friends to appear as another person. Using this fake identity, they may trick someone into believing they are in an online romance before asking the person to send money, gifts or nude images.

3 Conclusion

The Typology of Online Harms provides a comprehensive framework for understanding and categorizing various types of online harm through a human rights lens.

The typology plays a key role in identifying online harms and providing a foundational terminology for multistakeholder discussions. These discussions, in turn, can facilitate the creation of policies and interventions that effectively address online harms and reduce the associated risks.

This typology recognizes the complex nature of online safety, by classifying the threats into content, contact and conduct risks. Online harms can occur throughout the production, distribution and consumption of content (content) but can also arise as a result of online interactions with others (contact) and through behaviour facilitated by technology (conduct).

Furthermore, the typology categorizes these online harms. For example, it refers to threats to personal and community safety, such as child sexual exploitation material, pro-terror material and extremist context, among other types. Additionally, the typology identifies harms to health and well-being caused by content that promotes suicide or disordered eating. It also acknowledges the importance of dignity and privacy by including examples of bullying and harassment, doxxing and image-based abuse as violations of these principles. Deception and manipulation form another category within the typology, focusing on online harms related to disinformation and deceptive synthetic media.

The typology recognizes regional and local distinctions in how specific harms are defined and categorized in different jurisdictions. In this sense, it does not prescribe specific actions or assign severity ratings to online harms. Instead, it aims to serve as a valuable resource for companies to understand the online harms that may occur on their platforms.

While this typology offers a foundation for understanding online harms, new technologies, platforms and trends such as the metaverse, Web3 and generative AI may give rise to new forms of harm or exacerbate existing ones. Although future harms are not in the scope of this publication, the Coalition will consider creating a conceptual framework to ensure the approach to online harms is future-proof.

In conclusion, the Typology of Online Harms provides a comprehensive framework that contributes to global efforts to advance digital safety. By understanding different types of online harm, stakeholders can work collaboratively to develop effective policies, interventions and innovations that promote a safer digital ecosystem while respecting human rights and fostering positive online behaviours.

Appendix: Resources

Australian Government eSafety Commissioner, *Assessment Tools*: https://www.esafety.gov.au/industry/safety-by-design/assessment-tools.

Australian Government eSafety Commissioner, *Glossary of Terms*: https://www.esafety.gov.au/about-us/glossary.

Australian Government eSafety Commissioner, *Online Safety Act 2021: Abhorrent Violent Conduct Powers – Regulatory Guidance*, 2021: https://www.esafety.gov.au/sites/default/files/2022-03/Abhorrent%20Violent%20Conduct%20Powers%20Regulatory%20Guidance.pdf.

Berkman Center for Internet & Society at Harvard, *Enhancing Child Safety & Online Technologies: Final Report of the Internet Safety Technical Task Force to the Multi-State Working Group on Social Networking of State Attorneys General of the United States*, 2008: https://www.ojp.gov/ncjrs/virtual-library/abstracts/enhancing-child-safety-and-online-technologies-final-report.

Child Dignity Alliance, *Child Dignity in the Digital World: Technology Working Group Report*: https://www.childdignity.com/technical-working-group-report.

Digital Trust & Safety Partnership, *Trust & Safety Glossary of Terms*, 2023: https://dtspartnership.org/wp-content/uploads/2023/01/DTSP_Trust-Safety-Glossary13023.pdf.

DQ Institute, *Outsmart the Cyber-Pandemic: Empower Every Child with Digital Intelligence by 2020*, 2018 DQ Impact Report: https://www.dqinstitute.org/2018dq_impact_report/.

ECPAT International, *Luxembourg Guidelines*: https://ecpat.org/luxembourg-guidelines/.

EU Disinfo Lab, *Gender-Based Disinformation: Advancing Our Understanding and Response*, 20 October 2021: https://www.disinfo.eu/publications/gender-based-disinformation-advancing-our-understanding-and-response/.

European Parliament, Policy Department for Structural and Cohesion Policies, Brussels, *Research for CULT Committee – Child Safety Online: Definition of the Problem*, 2018: https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/602016/IPOL_IDA(2018)602016_EN.pdf.

Europol, *Internet Organised Crime Threat Assessment*, 2018: https://www.europol.europa.eu/internet-organised-crime-threat-assessment-2018.

IWF, *Internet Watch Foundation Annual Report 2022*: www.iwf.org.uk.

Kijkwijzer, *Youth Protection Roundtable Toolkit* (in Dutch): https://www.kijkwijzer.nl/nieuws/nicam-dossier-5-naar-een-safer-internet-gepubliceerd/.

Livingstone, S. and M. Stoilova, *4 Cs of Online Risk: Short Report & Blog on Updating the Typology of Online Risks to Include Content, Contact, Conduct, Contract Risks*, Children Online: Research and Evidence, 2021: https://core-evidence.eu/posts/4-cs-of-online-risk.

Meta, *Facebook Community Standards*: https://transparency.fb.com/en-gb/policies/community-standards.

Ofcom, *Addressing Harmful Online Content: A Perspective from Broadcasting and On-Demand Standards Regulation*, 2018: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/120991/Addressing-harmful-online-content.pdf.

Reddit, *Reddit Content Policy*: https://www.redditinc.com/policies/content-policy.

Teimouri, M., S.R. Benrazavi, M.D. Griffiths and S. Hassan, *A Model of Online Protection to Reduce Children's Risk Exposure: Empirical Evidence from Asia*, Sexuality & Culture, vol. 22, issue 4, 2018, pp. 1205–1229: https://www.researchgate.net/publication/324808035_A_Model_of_Online_Protection_to_Reduce_Children%27s_Online_Risk_Exposure_Empirical_Evidence_From_Asia.

Thorn.org, *Child Pornography Is Sexual Abuse Material*: https://www.thorn.org/child-pornography-and-abuse-statistics/.

TikTok, *Community Guidelines*: https://www.tiktok.com/community-guidelines?lang=en.

Twitter Help Center, *Platform Use Guidelines*: https://help.twitter.com/en/rules-and-policies.

UNICEF, *Children in a Digital World: The State of the World's Children 2017*: https://www.unicef.org/reports/state-worlds-children-2017.

United Nations Officer of the High Commissioner, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment*, 2021: https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation.

WHO, *ICD-11: International Classification of Diseases 11th Revision*: https://icd.who.int/en.

Woodlock, D., *ReCharge: Women's Technology Safety, Legal Resources, Research & Training*, 2015: https://wesnet.org.au/wp-content/uploads/sites/3/2022/05/ReCharge-national-study-findings-2015.pdf.

World Economic Forum, *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*, 26 May 2023: https://www.weforum.org/reports/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies.

World Economic Forum, *Global Principles on Digital Safety: Translating International Human Rights for the Digital Context*, 9 January 2023: https://www.weforum.org/whitepapers/global-principles-on-digital-safety-translating-international-human-rights-for-the-digital-context.

YouTube Help: https://support.google.com/youtube#topic=9257498.

Contributors

World Economic Forum

Minos Bantourakis
Head, Media, Entertainment and Sport Industry

Agustina Callegari
Project Lead, Global Coalition for Digital Safety

Daniel Dobrygowski
Head, Governance and Trust

Cathy Li
Head, AI, Data and Metaverse, Centre for the Fourth Industrial Revolution; Member of the Executive Committee

Lead Authors

Daniel Child
Industry Affairs and Engagement Manager, eSafety Commissioner

John-Orr Hanna
Chief Intelligence Officer, Crisp, a Kroll Business

Adam Hildreth
Founder, Crisp, a Kroll Business

Julie Inman Grant
Commissioner, Australian eSafety Commissioner

Acknowledgements

Antigone Davis
Vice-President, Global Head of Safety, Meta

Julie Dawson
Chief Policy and Regulatory Officer, Yoti

Inbal Goldberger
Vice-President of Trust and Safety, ActiveFence

Susie Hargreaves
Chief Executive Officer, Internet Watch Foundation

Lisa Hayes
Head of Safety Public Policy and Senior Counsel, Americas, TikTok

Farah Lalani
Global Vice-President, Trust and Safety Policy, Teleperformance

Heidi Larson
Professor of Anthropology, Risk and Decision Science, London School of Hygiene and Tropical Medicine

Deepali Liberhan
Director, Safety Policy, Global (Head, Regional and Regulatory), Meta

Susan Ness
Distinguished Fellow, Annenberg Public Policy Center of the University of Pennsylvania

Akash Pugalia
Global President, Trust and Safety, Teleperformance

Katherine Sandell
Global Head of Platform Risk Trust and Safety, Google

Ian Stevenson
Chief Executive Officer, Cyacomb

David Sullivan
Executive Director, Digital Trust and Safety Partnership

John Tanagho
Executive Director, International Justice Mission Center to End Online Sexual Exploitation of Children

David Wright
Director, UK Safer Internet Centre

Endnotes

1. World Economic Forum, *Global Principles on Digital Safety: Translating International Human Rights for the Digital Context*, 9 January 2023: https://www.weforum.org/whitepapers/global-principles-on-digital-safety-translating-international-human-rights-for-the-digital-context.
2. World Economic Forum, *A Global Coalition for Digital Safety*, 2023: https://initiatives.weforum.org/global-coalition-for-digital-safety/home.
3. World Economic Forum, *Global Principles on Digital Safety: Translating International Human Rights for the Digital Context*, 9 January 2023: https://www.weforum.org/whitepapers/global-principles-on-digital-safety-translating-international-human-rights-for-the-digital-context.
4. World Economic Forum, *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*, 26 May 2023: https://www.weforum.org/reports/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies.
5. World Economic Forum, *Global Principles on Digital Safety: Translating International Human Rights for the Digital Context*, 9 January 2023: https://www.weforum.org/whitepapers/global-principles-on-digital-safety-translating-international-human-rights-for-the-digital-context.
6. Ibid.
7. Livingstone, Sonya, and Stoilova, Mariya, *The 4Cs: Classifying Online Risk to Children*, CO:RE Short Report Series on Key Topics, SSOAR, 2021: https://doi.org/10.21241/ssoar.71817.
8. United Nations Convention on the Rights of the Child, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment*, 2 March 2021: https://www.unicef.org/bulgaria/en/media/10596/file.
9. United Nations, *Universal Declaration of Human Rights*, 2015: https://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf.
10. The International Centre for Missing and Exploited Children, *Words Matter*: https://www.icmec.org/resources/terminology/.
11. Gilbert + Tobin Centre of Public Law, *Review of Commonwealth Laws for Consistency with Traditional Rights, Freedoms and Privileges*, 25 February 2015: https://www.alrc.gov.au/wp-content/uploads/2019/08/22._org_gilbert_and_tobin_centre_for_public_law_alrc_freedoms_sub.pdf.
12. United Nations Convention on the Rights of the Child, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment*, 2 March 2021: https://www.unicef.org/bulgaria/en/media/10596/file.
13. United Nations, *International Covenant on Economic, Social and Cultural Rights*, 16 December 1966: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights.
14. United Nations Convention on the Rights of the Child, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment*, 2 March 2021: https://www.unicef.org/bulgaria/en/media/10596/file.
15. United Nations, *International Covenant on Civil and Political Rights*, 16 December 1966: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights#:~:text=Article%2020,-1.&text=Any%20propaganda%20for%20war%20shall,shall%20be%20prohibited%20by%20law.
16. United Nations, *Universal Declaration of Human Rights*: https://www.un.org/en/about-us/universal-declaration-of-human-rights#:~:text=Article%201,in%20a%20spirit%20of%20brotherhood.
17. *Facilitator's Tips: Strategies to Overcome Challenges*: https://cdn.icmec.org/wp-content/uploads/2020/02/20213159/PasP-facilitator-tips.pdf (adapted from Prevent Child Abuse, Facilitator's Guide to Resilience and Appendix D p. 18 from Dealing with Difficult Audience types from Prevent Child Abuse America: https://preventchildabuse.org/resources/resilience/).
18. United Nations Convention on the Rights of the Child, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment*, 2 March 2021: https://www.unicef.org/bulgaria/en/media/10596/file.
19. Ibid.

WORLD ECONOMIC FORUM

COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org