

Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies

INSIGHT REPORT

MAY 2023

Contents

Foreword	3
Executive summary	4
Introduction	5
1 Risk assessment framework	7
2 Bank of case studies	9
Case study 1 Trust and safety best practices – DTSP framework	12
Case study 2 Human Rights Due Diligence (HRDD) – GNI assessment	19
Case study 3 Systems/outcomes-based approach – New Zealand Code of Practice	25
Case study 4 Safety by design – The Australian eSafety Commissioner’s Safety by Design start-up assessment tool	31
Case study 5 Child safety – gaming, immersive worlds and the metaverse	36
Case study 6 Algorithms – AI impact assessment tool	42
Conclusion	46
Contributors	47
Endnotes	49

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.

© 2023 World Economic Forum. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, including photocopying and recording, or by any information storage and retrieval system.

Foreword



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



Cathy Li
Head, AI, Data and Metaverse;
Centre for the Fourth Industrial
Revolution; Member of the
ExCom, World Economic Forum



David Sullivan
Executive Director, Digital Trust
& Safety Partnership



Gill Whitehead
Online Safety Group
Director, UK Office of
Communications (Ofcom)

The digital world has profoundly impacted our lives, revolutionizing how we live and work. It has enabled us to connect with people from all corners of the globe, expand our knowledge horizons and drive innovation to unprecedented levels. However, as with any new frontier, the digital landscape has also presented various challenges, particularly concerning harmful content and online behaviour.

Recognizing the need to address these challenges, the World Economic Forum's [Global Coalition for Digital Safety](#) has brought together a diverse group of leaders to accelerate public-private cooperation to tackle harmful content and conduct online. Central to the coalition's efforts was the development of the [Global Principles on Digital Safety](#), emphasizing

the importance of an effective risk management framework to help organizations proactively foster a safer online environment.

To this end, the coalition is focused on developing a cross-jurisdictional baseline framework for understanding and assessing digital safety risks. This paper represents the first output of this collaborative effort, offering a comprehensive risk assessment framework accompanied by a bank of case studies that demonstrate the framework's practical application. This work is the result of extensive discussions engaging a broad range of stakeholders across sectors and jurisdictions, making it an invaluable tool for anyone interested in the intersection of technology, policy and human rights.

Executive summary

The [Global Coalition for Digital Safety](#) is a unique public-private platform that brings together [key stakeholders](#) to tackle the issue of harmful content and conduct online. The coalition recognizes that the digital world has brought unprecedented opportunities for people worldwide to connect, learn and innovate. However, it also acknowledges that this new world has its share of challenges, particularly regarding digital safety. To address these challenges, the coalition has embarked on three workstreams: 1) the [Global Principles on Digital Safety](#), addressing the critical question of how international human rights principles translate into a digital context, aiming to advance digital safety in a rights-respecting way, drive multistakeholder alignment and enable positive behaviours and actions across the digital ecosystem, 2) a toolkit for digital safety design interventions and innovation, developing a “typology of online harms” aimed at facilitating cross-platform and cross-jurisdictional discussions, and identifying what technology, policy, processes and design interventions are needed to advance digital safety, and 3) a risk assessment framework, aiming to develop a cross-jurisdictional baseline framework for understanding and assessing digital safety risks.

This paper is the first output of the third workstream, providing a **risk assessment framework** accompanied by a **bank of case studies** demonstrating how the framework might be applied in practice.

This work takes place in an evolving landscape: until recently, organizations were undertaking digital safety risk assessments on a voluntary basis, now a growing number of regulations are being proposed

and enacted provisioning risk assessments.

This **framework** draws on existing human rights frameworks, enterprise risk management best practices and evolving regulatory requirements to clarify the factors that should be used to clarify digital safety risks and sets out a methodology for how stakeholders should assess these risk factors in the digital ecosystem. It proposes a holistic approach that links risks and realized harms – or adverse impacts – in a cyclical process, ultimately leading to a virtuous circle and driving continuous improvement. It is harm and service-agnostic, aiming to serve all stakeholders.

The **case studies** highlight the variety and interconnectedness of existing risk assessment frameworks and approaches while substantiating the complexity of the subject matter, providing an overview of how existing frameworks are designed and leveraged and how a risk assessment framework can be applied in practice to a specific technology, type of harm or type of service.

The paper will be complemented by three forthcoming publications: 1) *Typology of Online Harms*: classification of online harms categories providing a common foundational language for multilateral stakeholder discussions, 2) *Risk Factors, Metrics and Measurement*: identification of characteristics that could contribute to adverse impacts, and metrics or measurement approaches that could be considered as part of risk assessments, and 3) *Solution-Based Interventions*: supporting companies steering towards more effective digital risk identification and reduction, harm prevention, mitigation and repair, drawing on Safety by Design principles and trust and safety best practices.

Introduction

Effective risk management can help organizations be more proactive and effective in fostering a safer online environment.



“ The relationship between specific aspects of a product, service or policy and the harms they might cause or contribute to can be difficult to assess or predict.

Digital services are at the heart of economic, educational, social and political affairs across the globe. They have propelled economic growth and innovation in countries worldwide while playing a critical role in enabling and empowering individuals to enjoy their human rights. However, the fast-paced development of new technologies and the immense volume of online activity also generate continuous risks to people, communities and societies. As the way people use technologies continues to evolve, so will the harms associated with online content and behaviour and the potential measures to address them. In this context, focusing on effective risk management can help organizations be more proactive and effective in fostering a safer online environment.

Yet digital safety risk assessment remains a nascent and evolving discipline. Digital safety requires a complicated range of deliberations, balancing legal, policy, ethical, social and technical considerations. Similar complexities apply to the nature of harms, which can be highly local or context-specific and differ across different communities, countries or regions. Finally, the relationship between specific aspects of a product, service or policy and the harms they might cause or contribute to can be difficult to assess or predict.

A range of frameworks, assessment methodologies and operational guidelines already exist to help manage complex risks. For example, the United Nations Guiding Principles on Business and Human Rights (UNGPs) sets out a human rights due diligence process that includes assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses and communicating how impacts are addressed. Enterprise risk management processes, while typically focused mainly on a company's business interests rather than broader digital safety objectives, can similarly serve as a useful tool or starting point.

Until recently, organizations that undertook digital safety risk assessments did so voluntarily, developing and adopting voluntary frameworks such as the Digital Trust & Safety Partnership or the Aotearoa New Zealand Code of Practice for Online Safety and Harms. That is now changing, in the context of a growing number of regulatory regimes that have been proposed or enacted, which include provisions around risk assessments. These can be broad, such as the systemic risk assessments under the EU's Digital Services Act, or focus on specific rights (e.g. data protection impact assessments), technologies (e.g. the EU Artificial Intelligence Act) or vulnerable groups (e.g. the UK Age-Appropriate Design Code). Requirements around dedicated online safety risk assessments also appear in the Australian Online Safety Act, Singaporean Online Safety Bill and the proposed UK Online Safety Bill, among others.

As a result, more organizations are developing and implementing digital safety risk management systems, aligning with the commitments in the [Global Principles on Digital Safety](#) “Embracing

innovative, evidence- and risk-based approaches to digital safety; for instance, through undertaking risk assessments”. This change often involves a mindset shift to prioritize safety and subsequent efforts to realign company values, management priorities and business incentives around this objective. Effective risk management relies on a culture of awareness and coordinated action across an organization, thus, these changes can be extensive and require significant resources. Yet the benefits are clear, from compliance with applicable regulatory regimes and effective mitigation of harms to access to capital and increased user engagement and retention.

This digital safety risk assessment framework draws on regulatory requirements and existing best practices to provide a high-level framework for understanding and assessing digital safety risks. It proposes a holistic approach that conceptually links risks – the potential for adverse impacts – and realized harms in a cyclical process. It is harm- and service-agnostic and can be leveraged by organizations of different sizes, scales and maturity levels. This holistic approach is intended to allow companies and stakeholders to adopt a more consistent approach to digital risk assessment while encouraging actors to assess and address safety risks in the round, encompassing the potential harm to both users and non-users and the impact across different human rights including safety, access to information, freedom of expression and privacy, among others. An overview of the online harms in the scope of risk assessments will be provided in the forthcoming *Typology of Online Harms* paper.

Accompanied by a range of case studies that illustrate the range of ways it can be operationalized in practice, the risk assessment framework is intended for use by stakeholders. This includes online service providers, safety tech and risk intelligence players, or content moderation and service providers, as well as the public sector (governments, regulators and international organizations), civil society (non-governmental organizations (NGOs), educators, youth) and investors (venture capitalists (VCs), start-ups, founders). It will be complemented by three forthcoming publications:

- **Typology of Online Harms:** classification and definition of online harms categories providing a common foundational language for multilateral stakeholder discussions.
- **Risk Factors, Metrics and Measurement:** identification of characteristics that could contribute to adverse impacts (e.g. service functionalities, user base or business models) and metrics or measurement approaches that could be considered part of risk assessments.
- **Solution-Based Interventions:** solutions-focused interventions to support companies steering towards more effective digital risk identification, harm prevention, mitigation and repair, drawing on Safety by Design principles and trust and safety best practices.

1

Risk assessment framework

Organizations should address safety risks comprehensively, considering the impact on users, non-users and various human rights.

This risk assessment framework is the product of a multistakeholder group and is meant to serve as a baseline framework to structure the approach and discussions on digital safety risk assessments.

FIGURE 1 Risk assessment framework



0 Identify risk

How are **risk factors identified, categorized and prioritized** (including based on type of service, user base, geo-location and data storage)?



1 Reduce risk

What **policies, safety mechanisms and proactive workflows** are implemented to reduce the risk of harm from happening or proliferating?



2 Mitigate harm

What **mechanisms** are in place to **report** or **detect** harm, and what **decision-making frameworks and enforcement workflows** are implemented to mitigate it?



3 Repair harm

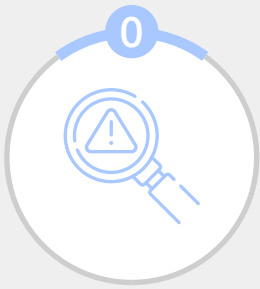
What mechanisms exist to **repair harm** (including post-incident response management and appeals mechanisms)?



4 Report

What mechanisms are implemented to measure and **monitor harm** on an ongoing basis and **fix systemic problems**? What **metrics** are recorded? How are reports used to **drive change**?

FIGURE 2 | Overall framework for the workstream approach – life cycle approach



Identify risk

Core assessment

How are **risk factors identified, categorized and prioritized**?

Key questions: examples

- How are the main risks associated with the service and user base identified?
- How are policies and associated user agreements (terms of service, community standards, etc.) defined and set out?
- How are users informed about policies and changes?
- Are policies consistent with the service's role in the technology stack?
- How are inherent risks related to your business and operating models evaluated?
- How are the likelihood, prevalence, scale, severity and impact of potential harms on the platform/service evaluated?
- How are diverse external stakeholders engaged in identifying risk?



Reduce risk

Core assessment

What **design decisions or safety mechanisms** are embedded in the platform to **reduce the risk of harm**?

Key questions: examples

- How are design decisions or system implementations undertaken to reduce the risk of harm from happening or proliferating?
- How are automated systems implemented to detect problematic content or conduct?
- How are proactive processes (e.g. pre-launch risk assessments, live experiments, third-party audits) put in place to reduce risk?
- What approaches and processes are taken to address the distinct safeguarding needs of minors and marginalized groups?
- Are age and identity verification mechanisms balanced with mechanisms for anonymity and free expression?
- How is personal data managed in terms of storage, sharing with third parties or use for commercial purposes like ads or recommendations?
- How are resources allocated to different geographies and languages?



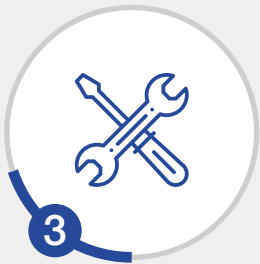
Mitigate harm

Core assessment

What mechanisms are in place to **detect harmful content or conduct** in the platform and perform relevant actions upon it?

Key questions: examples

- What processes are in place for the detection and removal/suspension/gating of violative content and/or for the identification of violative conduct?
- How is harm prevented from proliferating on the platform?
- What mitigation approaches are in place to reduce exposure to harm?



Repair harm

Core assessment

What mechanisms are implemented to **repair harm**?

Key questions: examples

- How are severe or systematic problems identified, escalated and prioritized?
- What processes are available to address complaints/appeals, and which provide avenues for support and guidance?
- How are affected stakeholders consulted to ensure that harms are mitigated/repaired?



Report

Core assessment

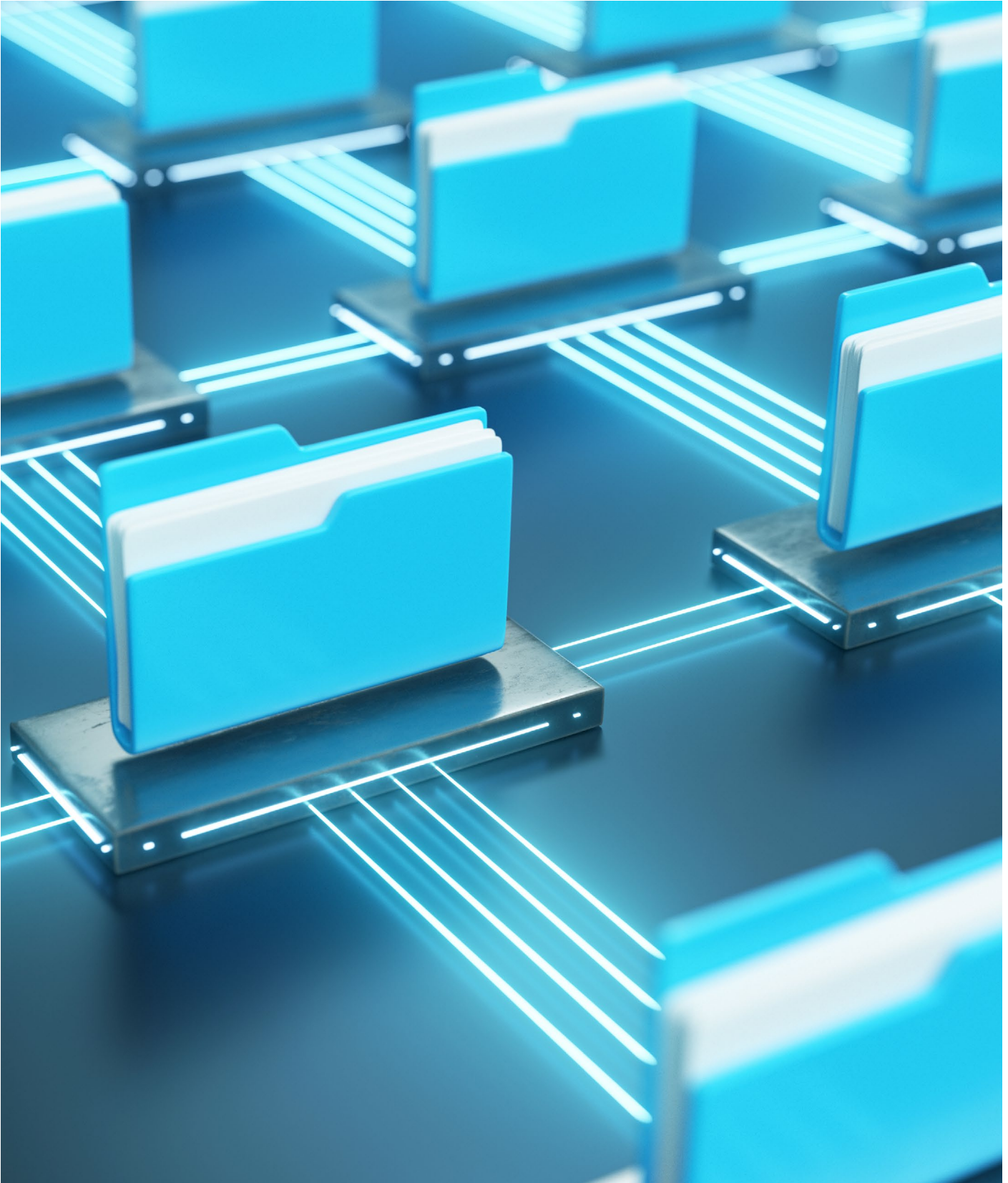
What reporting mechanisms are in place to **monitor impact** of harm and **fix systemic problems** that allow harm to reoccur?

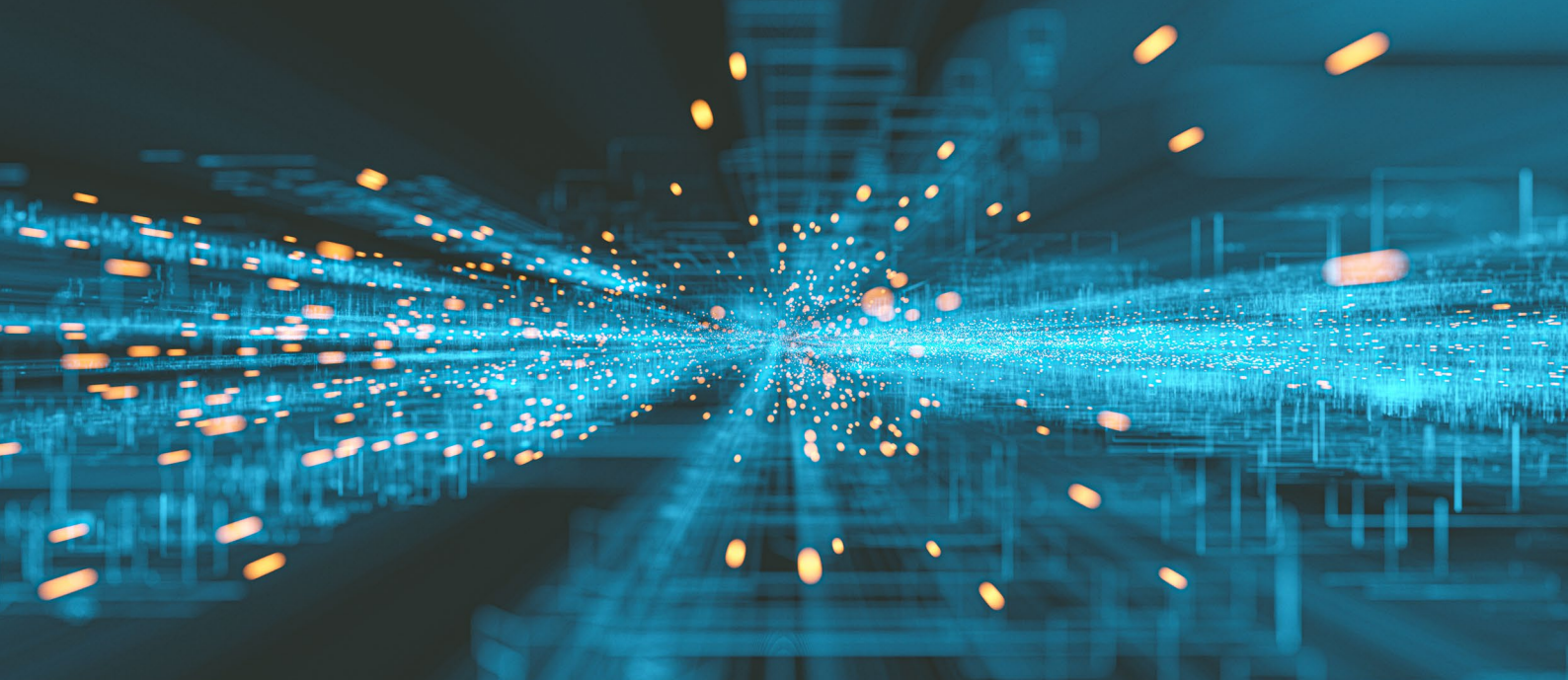
Key questions: examples

- What transparency and accountability measures are in place? Are internal and public reports made available?
- What key performance indicators (KPIs) and success metrics are used to measure the effectiveness of mitigations (e.g. accuracy, turn-around time)? How are these used to guide actions around detection and interventions?
- What processes are in place to drive product and process improvements? Can these be used to identify and address systemic problems?

2 Bank of case studies

Various approaches can be used to drive digital safety risk assessments and emphasize their interconnected nature.





The case studies aim to provide relevant support to all stakeholders engaged in online safety, including online service providers, safety tech and risk intelligence players, content moderation and service providers, the public sector (governments, regulators and international organizations), civil society (NGOs, educators, youth) and investors (VCs, start-ups, founders). They showcase the

wide array of potential approaches that can be undertaken to drive digital safety risk assessments, highlighting their interconnectedness. The first case studies (1, 2, 3 and 4) provide an overview of how existing frameworks are designed and leveraged, while the last two (5 and 6) are focused on how a risk assessment framework can be applied in practice to a specific technology, type of harm or type of service.



1. Trust and safety best practices – Digital Trust & Safety Partnership (DTSP) framework

Offers a content-agnostic framework of best practices that companies can use to address content- and conduct-related risks. It has the ambition to be used by online service providers of all sizes and types and requires taking an approach proportional to their scale and impact. To inform the level of depth of the assessment and

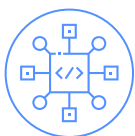
the consequent resource investment, it includes a “tailoring framework”: a proportionate, risk-based approach to determine the assessment level by evaluating organizational size/scale and potential impact, and a “maturity model”, building on experiences in other disciplines (e.g. software development, privacy, security).



2. Human Rights Due Diligence – Global Network Initiative (GNI) assessment

Centres on Human Rights Due Diligence (HRDD) for ongoing assessment, action, tracking and reporting to help tech companies respect freedom of expression and privacy when responding to government demands, pressures and restrictions. The role of independent third-party organizations

is of high interest in this case study: GNI’s multistakeholder board reviews periodic reports submitted by independent third-party assessors and evaluates GNI companies’ progress made in implementing the GNI commitments with improvement over time.



3. Systems/outcomes-based approach – New Zealand Code of Practice

A voluntary industry code and best practice self-regulatory framework, including a set of principles and commitments. It is a relevant example of a code tailored for a single local context (New Zealand) with its cultural and context specificities and provides transparency about efforts undertaken

to advance digital safety in the locale. It also shows the risk associated with developing differentiated country-specific codes, as this entails customization of the approach at a single country level, potentially generating inconsistencies across different locales and with principles of a global free and open internet.



4. Safety by design – the Australian eSafety Commissioner’s Safety by Design for start-ups assessment tool

The Safety by Design assessment tool provides a structured framework to bake user safety in and mitigate risks ahead of time in the design, development and deployment of online products and services. Risks are identified through a questionnaire, and practical tools and guidance materials, including templates, workflows and case studies of good practice, are provided to educate,

enhance capability, reduce risk and address safety gaps. This can include business model canvases, reporting mechanisms, content moderation workflow, product development processes and videos from tech industry experts – from leadership through to product developers and the trust and safety team. The toolkit is pragmatic and tailored for different maturities of companies, providing easy-to-use resources.



5. Child safety – gaming, immersive worlds and the metaverse

Focused on a virtual reality (VR)/metaverse gaming experience and child safety and child sexual abuse material (CSAM)-related risks, assessing risks along the end-to-end user experience concerning each aspect across user registration, payment methods,

commercial model, player interactions and many others. Interventions are designed for the specific user experience, showcasing that one-size-fits-all solutions do not work and that specific interventions are needed.



6. Algorithms – artificial intelligence (AI) impact assessment tool

Focused on the risks related to a search engine when combatting the spread of undesirable content. It analyses the following situations: 1) searching data voids (available relevant data is limited, non-existent or deeply problematic), and 2) when an unexpected event generates a lot of time-sensitive problematic

content. It showcases the challenge of a highly complex and unpredictable environment, with unpredictable “black swans” emerging, requiring set-up processes to identify spikes/sudden events and counteract them at speed.





CASE STUDY 1

Trust and safety best practices – DTSP framework



General information

1. High-level description of the case study

The DTSP¹ is an industry initiative dedicated to developing best practices, verified through internal and independent third-party assessments, to ensure consumer trust and safety when using digital services. DTSP brings together technology companies providing a wide range of digital products and services around a common approach to increasing trust and safety across the internet. All participating companies commit to the DTSP best practices framework (BPF), a content-agnostic tool that companies can use to address content- and conduct-related risks. The BPF consists of commitments to five fundamental areas of best practice: product development, governance, enforcement, improvement and transparency.

These commitments are underpinned by 35 specific trust and safety best practices spanning the phases of the risk assessment framework, from identification to reporting. It also includes concrete (but non-exhaustive) examples of the variety of activities and processes that organizations may have in place to mitigate risks from harmful content and conduct as appropriate to their individual product offerings and risk profiles. As the DTSP BPF is technologically and content agnostic, not all practices will apply to all products. Specific practices related to risk assessment include:

- Under product development: “Use in-house or third-party teams to conduct risk assessments to better understand potential risks”.
- Under improvement: “Use risk assessments to determine the allocation of resources for emerging content- and conduct-related risks”.

2. Context and main goals of the case study

In 2022, 10 DTSP partner companies conducted internal assessments of their implementation of the DTSP BPF using the organization’s assessment methodology, [The Safe Framework](#). The goal of these assessments was to help organizations understand how their trust and safety practices are working and how they support their adherence to the DTSP BPF. This case study summarizes the outcomes and key findings from this initial assessment of how it is implemented in practice. It’s worth noting that while these assessments complement broader efforts to assess the risks of products and services, it is not a product risk assessment tool and does not replace such efforts. The BPF also focuses narrowly on content- and conduct-related risks, so it is not designed to cover risks in other areas, such as security and privacy.

FIGURE 3 DTSP inventory of 35 best practices

Product development	Product governance	Product enforcement	Product governance	Product transparency
<ol style="list-style-type: none"> 1. Abuse pattern analysis 2. Trust and safety consultation 3. Accountability 4. Feature evaluation 5. Risk assessment 6. Pre-launch feedback 7. Post-launch evaluation 8. User feedback 9. User controls 	<ol style="list-style-type: none"> 1. Policies and standards 2. User-focused product management 3. Community guidelines/rules 4. User input 5. External consultation 6. Document interpretation 7. Community self regulation 	<ol style="list-style-type: none"> 1.1. Roles and terms 1.2. Operational infrastructure 1.3. Tooling 2. Training and awareness 3. Wellness and resilience 4. Advanced detection 5. User reporting 6.1. Enforcement prioritization 6.2. Appeals 6.3. External reporting 7. Flagging processes 8. Third parties 9. Industry partners 	<ol style="list-style-type: none"> 1. Effectiveness testing 2. Process alignment 3. Resource allocation 4. External collaboration 5. Remedy mechanisms 	<ol style="list-style-type: none"> 1. Transparency reports 2. Notice to users 3. Complaint intakes 4. Researcher and academic support 5. In-product indicators

Source: DTSP, *The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices*, 2022.

3. Target population(s) affected by the intervention and anticipated impacts

DTSP assessments explore how participating companies manage their content- and conduct-related risks. The target population includes all categories of users and is not limited to specific individuals, groups or categories of service users.

Existing methodologies/frameworks

4. Relevant existing risk methodologies and frameworks are taken into consideration

The DTSP best practices framework was inspired by the trajectory of cybersecurity and other tech disciplines, which have fostered more robust and consistent approaches to risk management by developing frameworks, assessments and standards. The development of the NIST Cybersecurity Framework and ISO 27000 standards, for example, has matured and organized the industry and enabled certification through audits and third-party assessments.

DTSP was also created with a view towards other enterprise risk management frameworks already used by companies, such as the Committee of Sponsoring Organizations of the Treadway Commission (COSO) framework for assessing internal controls related to financial reporting. It also aligns with the UN Guiding Principles on Business and Human Rights, especially in that the best practices seek to address content- and conduct-related risks to individuals' human rights, not just risks to the company.

Finally, DTSP is responsive to civil society initiatives, including the Santa Clara Principles on Transparency and Accountability in Content Moderation, which sets out commitments to transparency, notice and appeals that have been embedded into the best practices framework.

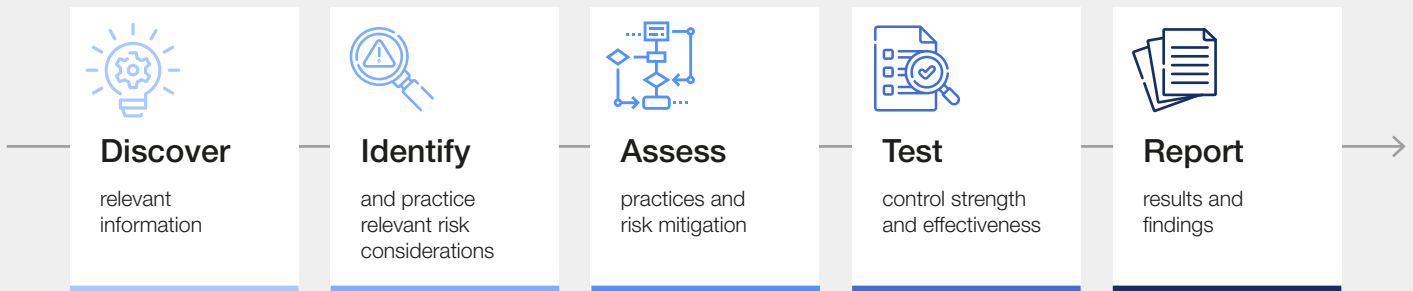
5. Measurement frameworks and assessment metrics involved

The Safe Framework uses a proportionate, risk-based approach to determine the depth of the assessment. Participating companies are assigned to one of three levels by evaluating objective factors for organizational size, scale and potential impact of the product or service being evaluated. Factors for evaluating organizational size and scale included annual revenue and the number of employees for the product or service in scope for assessment. Factors for product impact included user volume and product features that may implicate risk or complexity. This tailoring framework provides a common approach that companies with different resource levels can apply without imposing the same requirements on products with dissimilar functions, features or user base profiles. This helps ensure that assessment approaches are not overly resource-intensive in ways that would advantage those companies with the most resources available. DTSP's tailoring approach is described in detail in their 2021 report.²

After tailoring the assessment approach to the appropriate level, partner companies executed the safe assessments using a five-step methodology, from initial information-gathering or discovery to reporting results. A question bank was used to have a common resource as partners began the information discovery phase of the assessments.

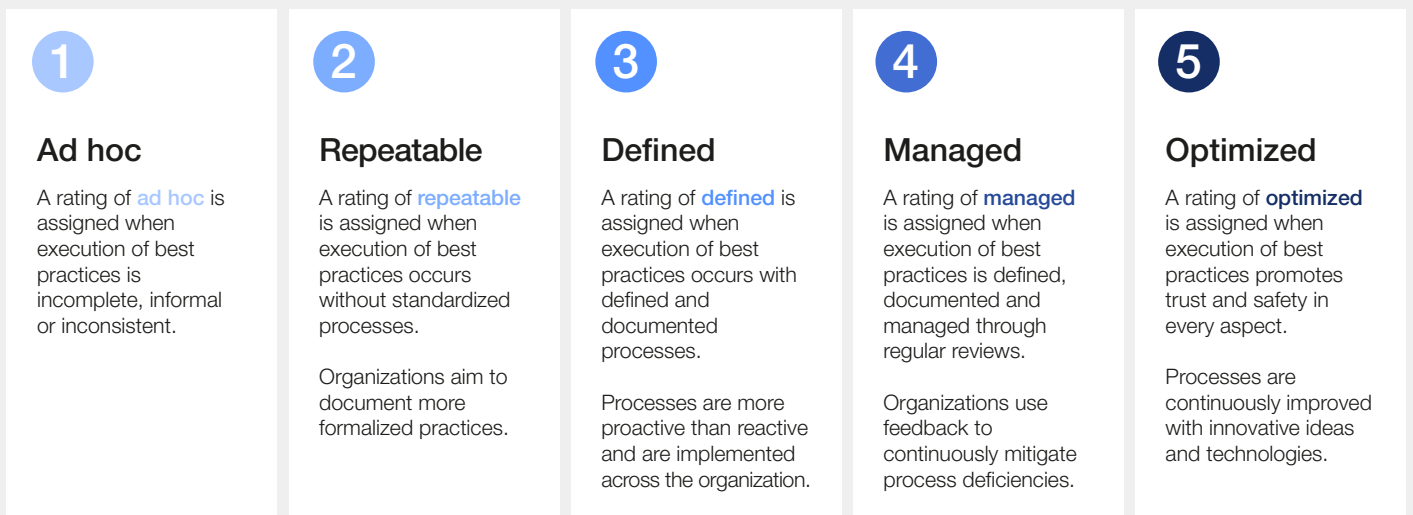
“ Assessment approaches are not overly resource-intensive in ways that would advantage those companies with the most resources available.

FIGURE 4 | Safe assessment – five-step methodology



Source: DTSP, *The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices*, 2022.

FIGURE 5 | DTSP maturity rating scale



Source: DTSP, *The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices*, 2022.

6. Legal or regulatory obligations that played a role in this case study

DTSP is a voluntary partnership that aims to develop best practices that can reduce the harms associated with online content and conduct. This can supplement and complement efforts to comply with certain regulatory requirements, particularly those related to risk assessment and audits. For example, under the EU Digital Services Act, very large online platforms and search engines have systemic risk assessment and audit requirements, but smaller platforms do not. The DTSP framework provides a proportionate means by which smaller platforms can begin to align their content risk management efforts with this evolving regulatory regime.

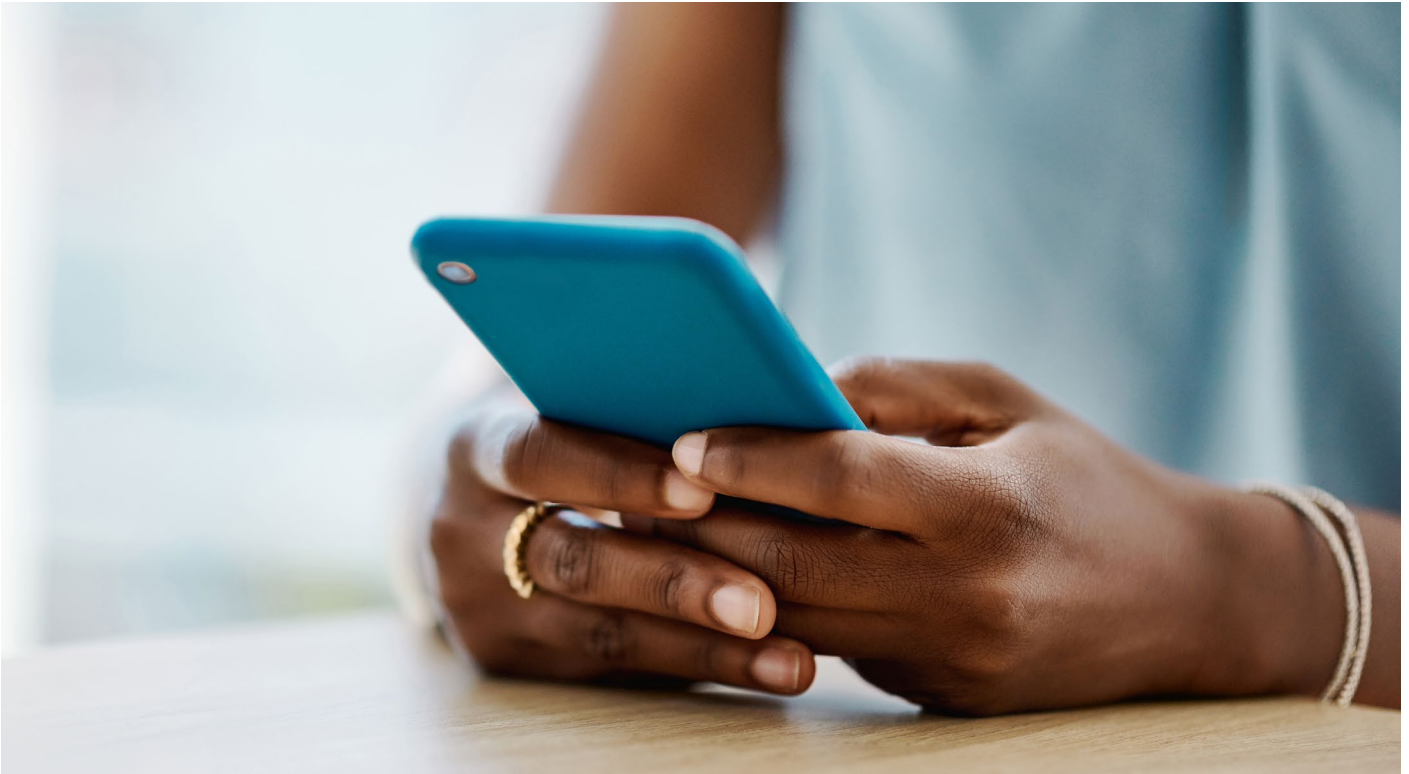
7. Benefits and risks associated with the approach taken

A key benefit of the DTSP approach is to provide an industry-wide guide to addressing content-and conduct-related risks that is adaptable to diverse products and services, as well as differing levels of organizational maturity. DTSP is also content and technologically agnostic, so companies facing very different risks can align around a common set of practices. However, this means that only some practices are suitable for some services. For example, the governance best practice on “community self-regulation”, which provides forums for community-led governance and tools for community moderation, might not apply to digital products and services that are not social in nature, such as a file storage service.

DTSP approached the inaugural safe assessments as a learning exercise where companies took individual approaches to scoping their assessments. Some companies assessed one or more products, while others assessed a central trust and safety function or a component of that function. Some companies focused on particular commitments and practices, while others assessed all of them.

The DTSP assessment methodology did present certain challenges and limitations, including:

- Comparability constraints from the assessment of different products and functions
- Consistency challenges from self-assessments and different approaches to implementation
- Applicability limitations relating to aggregating results for public reporting.



Implementation

8. Changes from the current state or practice that resulted from the risk assessment undertaken

The specific, company-by-company changes resulting from undertaking safe assessments are not made public. Instead, DTSP aggregates and anonymizes results to provide industry-level insight into trust and safety practices. Summary conclusions included:

Successes: many companies reported a mature state of development for core content moderation practices

The areas where trust and safety teams reported relatively mature practices included core practices and activities that fall squarely within their domain and can be implemented unilaterally to some degree. These include constituting the teams

responsible for content policies and developing public-facing policy descriptions, as well as developing enforcement infrastructures that span people, processes, and technology, and notifying users whose content is subject to enforcement action by the platform for violating its policies.

Areas for improvement: many of the least mature practices relate to user feedback and external collaboration

According to the self-assessments, three of the least mature practices are related to incorporating user and third-party perspectives into trust and safety policy and practices. This illustrates the internal focus of trust and safety functions. Until recently, trust and safety has developed with less external engagement outside of companies. The least mature of all assessed practices is the creation of processes to support academic and other researchers working on the relevant subject matter.

“ Trust and safety frameworks will formalize processes, improve the implementation of best practices, clarify who is accountable and encourage oversight.

Areas of ongoing improvement: integrating trust and safety into product development

Most assessments indicated that companies were in the process of formalizing the relationship between trust and safety and product teams to better integrate these perspectives into product development. Specifically, regarding risk assessment, the safe assessments showed that using risk assessment to drive resource allocation across emerging risks as part of the product improvement commitment lagged behind in using risk assessments in product development. Some assessments described the collaboration between trust and safety, policy and product teams to develop a methodology for ad-hoc risk assessments based on product launches and other key events but noted the need for more mature capabilities. Such capabilities include the performance of annual risk assessments to identify and report on top risk areas or the development of systemic risk strategies. This can support the kind of systemic risk assessments that will be required under some regulations, as well as public reporting on these activities.

9. Investment is required in terms of resources and timelines for implementation

Internal DTSP company assessments required sustained efforts by internal teams responsible for DTSP participation (often teams working on content policy, trust and safety policy, or related matters). Securing buy-in for this approach from senior management and trust and safety teams

and executing the assessments required up to several months of work. In addition, teams with expertise in risk and compliance and other assessment frameworks are key for testing controls, especially for the more intensive level two and level three assessments.

10. Other outcomes

Each DTSP company assessment identified areas of opportunity and development for the partner companies. Some examples include the planned rollout of new trust and safety governance and oversight frameworks. These frameworks will formalize processes, improve the implementation of best practices, clarify who is accountable for outline structures and encourage oversight. Assessments also indicated that those best practices, overlapping with requirements regulations, will receive special emphasis.

This process also generated useful feedback for DTSP: opportunities to review and refine best practices where there may be duplication or overlap; opportunities to identify and share innovative practices that could be added to the BPF; clarifying the maturity scale to improve objectivity; and developing shared tools and resources to help with future assessments.

Finally, opportunities were identified for external stakeholders to look to the BPF to understand how the industry is addressing content- and conduct-related risks, and for policy-makers to ensure that clear legal frameworks support company implementation of best practices.

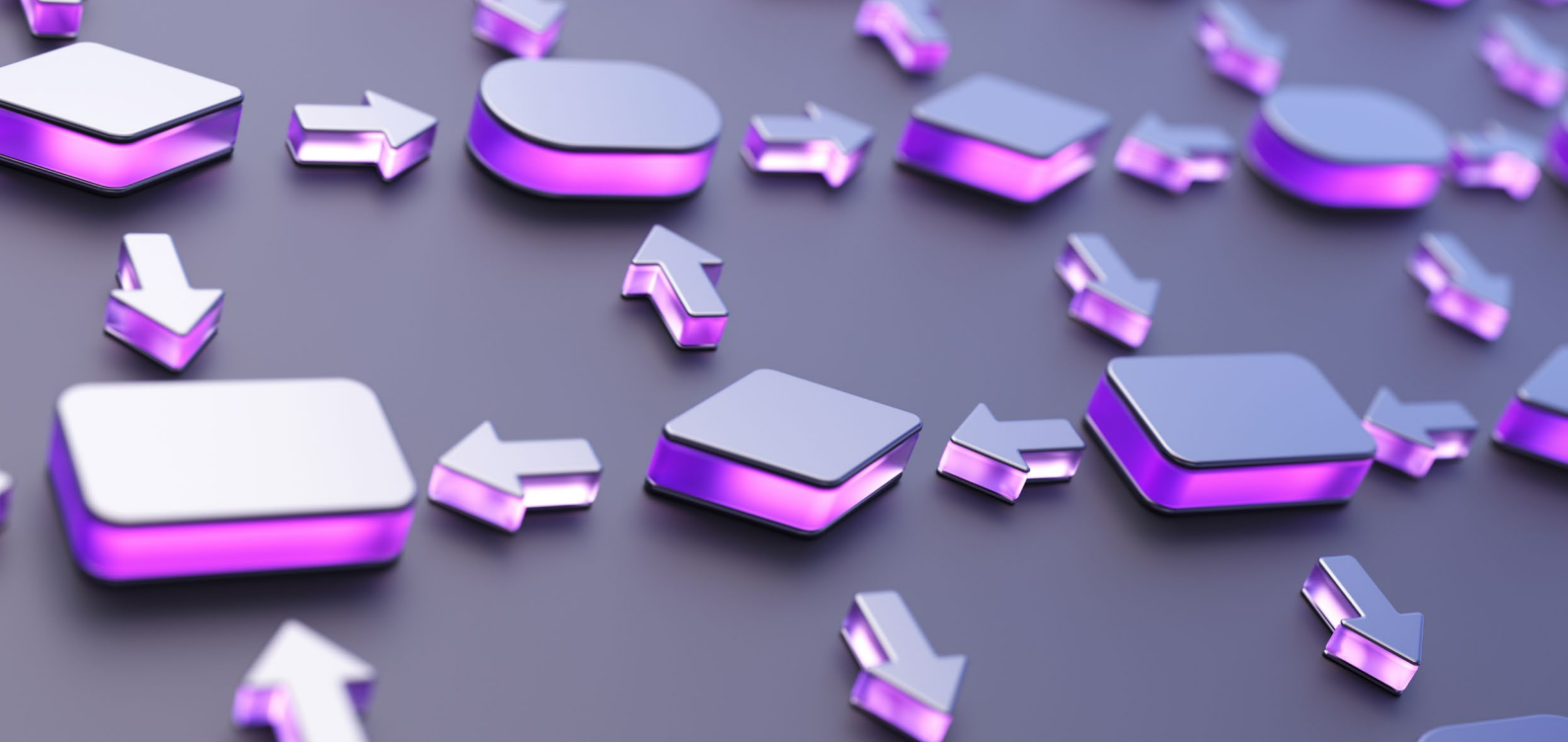




CASE STUDY 2

Human Rights Due Diligence (HRDD) – GNI assessment





General information

1. High-level description of the case study

The Global Network Initiative Principles on Freedom of Expression and Privacy and their accompanying implementation guidelines (together, the “GNI framework”) establish a specific framework to help tech companies respect freedom of expression and privacy when interacting with and responding to government demands, pressures and restrictions. Like broader (non-tech specific) frameworks (i.e. the Organisation for Economic Co-operation and Development (OECD) Guidelines for Multinational Enterprises and the UN Guiding Principles on Business and Human Rights), the GNI framework centres on HRDD as a framework for ongoing assessment, action, tracking and reporting of company efforts to identify and address human rights risks.

This case study examines how HRDD is described in the GNI framework, including its relationship to Human Rights Impact Assessments (HRIA) and broader company-wide HRDD efforts. The case study also describes how GNI members come together to collectively assess, learn from and improve company efforts through its independent assessment process. This assessment process is not a risk assessment, but reviews whether and how companies use the GNI Principles to embed HRDD in their policies, procedures and operations with a focus on the rights to privacy and freedom of expression.

“ HRDD starts by identifying potential human rights impacts and defines appropriate action to avoid, prevent and mitigate harm, including remedies for adverse impacts.

2. Context and main goals of the case study

HRDD includes all phases of the World Economic Forum’s risk assessment framework. HRDD starts by identifying potential human rights impacts – taking the list of internationally recognized human rights as a reference point and prioritizing the most salient – and defines appropriate action to avoid, prevent and mitigate harm, including remedies for adverse impacts. Core to effective HRDD is meaningful consultation with potentially affected stakeholders, and the GNI framework calls on companies to draw on a range of sources, including voices from inside relevant countries, human rights groups, government bodies and international organizations when assessing actual and potential human rights impacts. HRDD also involves sufficient communication with the public for the company’s approach to addressing human rights impacts to be effectively evaluated.

The GNI Principles state that member companies “will be held accountable through a system of (a) transparency with the public and (b) independent assessment and evaluation of the implementation of these Principles”⁴. Assessment reports, developed by independent, accredited assessors, include sensitive, non-public information illustrating how member companies are implementing the GNI framework, including with respect to HRDD. This includes both information about systems and processes, as well as selected case studies demonstrating how policies and procedures are implemented in practice. GNI’s multistakeholder board uses these reports to determine whether each company is implementing the framework in good faith with improvement over time.

3. Target population(s) affected by the intervention and anticipated impacts

GNI's multistakeholder board brings a range of perspectives and expertise on relevant regional and country-specific challenges, vulnerable groups and technological impacts to their engagement in this assessment process. A key target population during HRIA is often individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized. This is highly contextual – someone may be powerful in one context yet vulnerable in another – but can be considered across four dimensions:

- Formal discrimination: Laws or policies that favour one group over another.
- Societal discrimination: Cultural or social practices that marginalize some and favour others.
- Practical discrimination: Marginalization due to life circumstances, such as poverty.
- Hidden groups: People who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants and sexual assault victims.



Existing methodologies/frameworks

“ HRDD prioritizes circumstances where the risk of adverse impacts is most significant based on the criteria of scope, scale, remediability and likelihood.

4. Relevant existing risk methodologies and frameworks taken into consideration

The OECD guidelines and UNGPs set out a coherent and time-tested understanding of how companies can best comply with their responsibility to respect human rights. The OECD guidelines date to 1976 and were updated in 2012 in conformity with the UNGPs, which were unanimously endorsed by the UN Human Rights Council in 2011. The GNI framework was launched in 2008 and revised in 2017 to ensure consistency with these two broader approaches, building on them by zeroing in on two of the most salient rights for information and communications technology providers – privacy and freedom of expression – and addressing specific high-risk scenarios where tech companies' actions could impact these rights. Together, these frameworks have shaped decades of company practice, multistakeholder elaboration, experience and expertise.

Recent risk-focused regulatory efforts (including those targeting the tech sector specifically and those that apply to large companies regardless

of the sector) have taken a variety of approaches to defining key terms, methodologies and assessment approaches. Many efforts are building consciously on existing business and human rights understanding of HRDD, but sometimes ignoring or contradicting them. Technology companies are finding that HRDD assessment methodologies grounded in the UNGPs provide an excellent foundation for compliance with a growing range of regulatory requirements.

HRDD prioritizes circumstances where the risk of adverse impacts is most significant based on the criteria of scope (the number of people impacted), scale (the gravity of the impact), remediability (whether the impact can be made good) and likelihood (such as the frequency of impacts). A company's most salient human rights risks tend to be the focus of deeper or company-wide HRIAs instead of routine HRDD. In the GNI context, the focus is risks associated with government demands, pressures and restrictions relating to freedom of expression and privacy. The GNI's third-party assessors prioritize scrutiny of the business functions, lines of business and geographic areas that are material to companies' impacts on these rights.



5. Measurement frameworks and assessment metrics involved

Human rights risk is highly contextual and can be challenging to measure. However, some useful assessment metrics include:

- Scope: The volume of government demands for data or content restrictions and the portion complied with can indicate the scope of risk as it relates to government demands, pressures and restrictions.
- Likelihood: The prevalence of policy-violating content based on sampling techniques can indicate the probability of content policy violations (for example, see the Meta prevalence⁵ metric and the YouTube violative view rate⁶).
- Remendability: The number of successful user appeals can indicate a well-functioning grievance mechanism.

The *GNI Assessment Toolkit*⁷ builds on these general metrics by providing a framework for identifying the key systems and policies companies should have in place to uphold their GNI commitments. It also sets out guidance for independent third-party assessors to help them review and verify these commitments. While not

a “measurement framework”, the toolkit helps companies ensure that assessment reports are relatively consistent and comparable with other companies over time. Some relevant provisions of the toolkit include the following questions:

- “What processes or mechanisms does the company have to identify potential risks to freedom of expression and privacy that may be connected to each of the following: a) Products, including the development of new products or substantial changes in existing products? b) Markets, including an evaluation of relevant local laws and practices at the time of market entry or product sale, and as those laws and practices change over time? c) Acquisitions and partnerships where the company has operational control? d) Other business relationships?”
- “When the company’s routine due diligence surfaces human rights issues for analysis, mitigation and prevention, how does the company prioritize among those human rights issues?”
- “How does the company decide whether a detailed HRIA, rather than routine HRDD, is required to develop effective prevention and mitigation strategies? Please discuss in relation to both product- and market-based risks”.
- “How does the company conduct an HRIA? Please provide specific examples if helpful”.

6. Legal or regulatory obligations that played a role in this case study

The GNI framework is narrower than recent regulatory developments (focused on freedom of expression and privacy) and deeper (involves multistakeholder review). However, a common thread between the GNI framework and recent regulatory developments is the deployment of HRDD methodologies based on the UNGPs. This approach is particularly evident in regulation from the EU, which has been leading on many tech regulatory fronts. As detailed in a recent article by Business for Social Responsibility (BSR)⁶ examples include:

- The EU General Data Protection Regulation (GDPR) requires that companies undertake “data protection impact assessments” that consider not just privacy but impacts against all rights contained in the EU Charter of Fundamental Rights, prioritizing the most severe risk to “data subjects”.
- The EU Digital Services Act (DSA) requires a “systemic risk assessment” encompassing actual or foreseeable impacts on rights contained in the EU Charter of Fundamental Rights, emphasizing vulnerable users and offering scope, scale and remediability as potential prioritization criteria.
- The EU Artificial Intelligence Act will require a “conformity assessment” for higher-risk applications of AI and uses the EU Charter of Fundamental Rights as the basis for understanding and classifying risk.
- The EU Corporate Sustainability Reporting Directive will require that companies take a “double materiality” approach to disclosure, where the prioritization of matters that affect the economy, environment and people (“impact

materiality”) will be based on concepts of scope, scale and remediability drawn from the UNGPs.

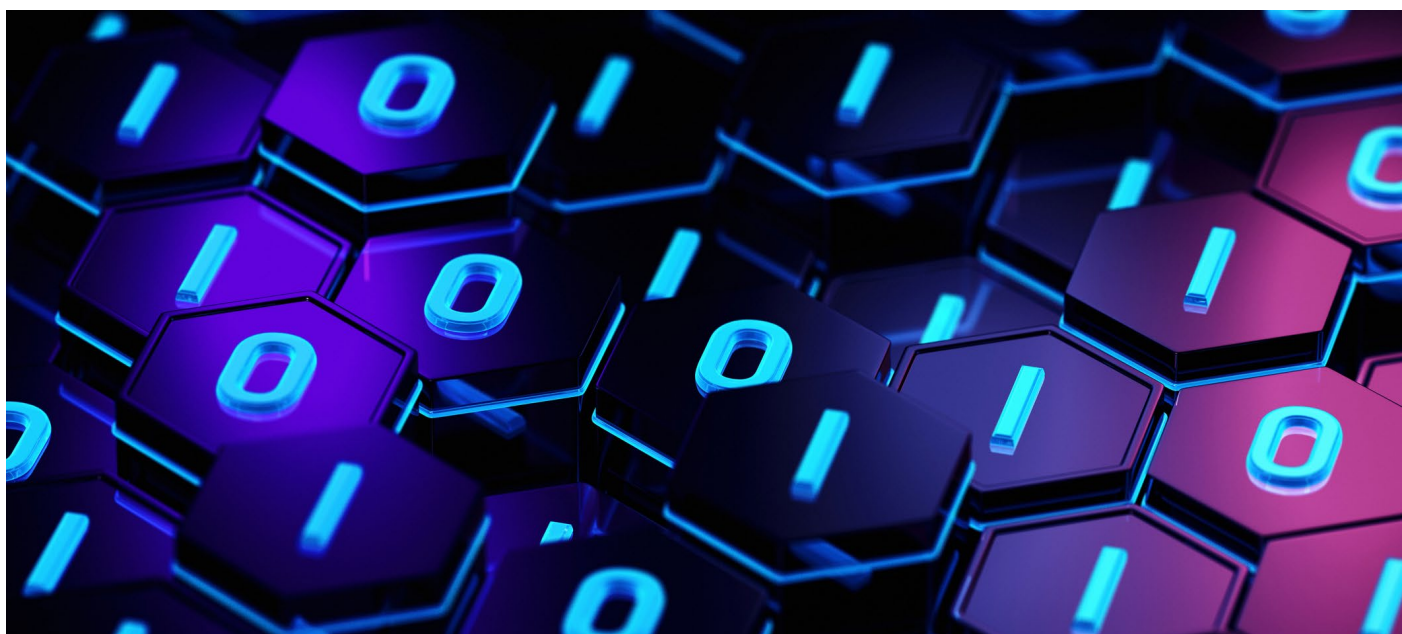
- The EU Corporate Sustainability Due Diligence Directive will establish a corporate due diligence duty, which will require identifying, preventing, mitigating and accounting for adverse human rights and environmental impacts and is likely to apply across companies’ entire value chains, including regarding the development and sale of products and services.

For companies, there are opportunities to consider the human rights-based synergy between these different requirements, such as connectivity through compliance processes, creating shared content across different assessments, or establishing an information architecture for reporting that positions these different disclosure requirements as an integrated whole.

For regulators, there is a need to maximize interoperability between these different requirements, including consistently emphasizing the relevance of all human rights, harmonizing criteria by which adverse impacts on people should be prioritized and creating more uniformity of disclosure requirements. Governments should aspire to model good practices for regulating digital content and conduct. The GNI framework details steps that companies can take to comply with local law and respect freedom of expression and privacy wherever they operate. GNI’s advocacy work has shown many governments putting in place laws and regulations that place undue pressure on companies to restrict access to content and services or share access to user data.

Preparing for compliance with applicable regulatory regimes will require significant, detailed and tailored activity to meet the requirements of each regulation. However, taking a consistent human rights-based approach based on the UNGPs will ease this process and enhance compliance with both the spirit and letter of each regulation.

“ Preparing for compliance with applicable regulatory regimes will require significant, detailed and tailored activity to meet the requirements of each regulation.



7. Benefits and risks associated with the approach taken

There are several benefits of a human rights-based approach to risk assessment, as detailed in the GNI framework and the UNGPs, including:

- Human rights, as enshrined in core UN treaties and declarations, are universal and provide a uniform approach to understanding the risks and impacts of business conduct.
- Centring risk on the people affected (“risks to people”) rather than the company (“business risks”).
- Requiring a methodical review against all internationally recognized human rights as a reference point since companies may potentially impact any of them.
- Ensuring that the interests of the most vulnerable are included as a matter of good process.
- The GNI assessment process is a unique accountability mechanism offering input from a diverse, multistakeholder board on companies HRDD and HRIA practices, including a review of non-public information.

There are several risks of human rights-based approaches to risk assessment, as detailed in the GNI framework and the UNGPs, that need to be addressed:

- Cumulative impacts may be missed when one event alone may not be a violation, but the same event repeated millions of times would be.
- Over-emphasis on a single company or type of company at the expense of the broader system. The recent work by the GNI and BSR to establish ecosystem-wide approaches to human rights due diligence (“Across the Stack Tool: Understanding HRDD under an Ecosystem Lens”⁹) is designed to address this risk.
- While the GNI framework is rooted in international human rights law and offers guidance on HRDD policies and practices taking stock of the full range of human rights, the GNI commitments are centred on the rights to freedom of expression and privacy.
- While human rights are universal, some countries take differing views of what they mean in practice.



Implementation

8. Changes from the current state or practice that resulted from the risk assessment undertaken

The GNI framework has driven a number of key advances in responsible business conduct over the last 15 years, including the development of robust company transparency reporting, the establishment of corporate-level human rights policies and board-level oversight of those policies, and increased conduct and transparency around the use of human rights impact assessments. The

growth of GNI's multistakeholder membership has also helped improve insight into and participation in corporate HRDD practice by a range of non-company stakeholders, including many majority world-based actors. While GNI's assessment focuses on freedom of expression and privacy risks at the intersection between companies and governments, the architecture of the GNI framework and GNI's multistakeholder structure help position companies to implement and demonstrate broader compliance with the holistic approach to HRDD set out in the OECD guidelines and UNGPs.



9. Investment required in terms of resources and the timeline for implementation

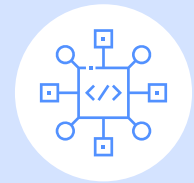
GNI assessments require a significant investment of time and resources on the part of companies GNI's staff and board. Employees of GNI companies responsible for leading assessments invest a lot of time in identifying relevant materials, facilitating interviews for assessors with key staff and explaining the GNI process internally. These same employees are often the ones responsible for the conduct of day-to-day HRDD. Non-company members of the multistakeholder GNI board carefully review each assessment report (the most recent assessment cycle covered 11 companies) and form study groups to identify questions to send to assessors and companies before each assessment review meeting. GNI staff are responsible for shepherding the entire process, including the organization of relevant meetings, training of assessors and producing the end-of-cycle public assessment reports.¹⁰

GNI conducts assessment reviews every three years. This timeline allows assessors and

companies time to conduct detailed and, at times, far-reaching analyses of underlying materials, including interviews with relevant staff. It also allows for a rigorous, post-facto review of the assessment process itself to produce improvements to the toolkit and assessor training. However, this timeline also means that information reviewed may become stale, which is especially concerning with regard to case studies.

10. Other outcomes

There are considerable concerns about the significant harms that are taking place on online services and how existing and emerging legal and regulatory frameworks for digital content regulation could lead to unintended human rights consequences and further fragmentation of the internet. Content regulation approaches can and should build on existing good practices and lessons learned around tech company implementation of HRDD, including the importance of conducting holistic due diligence and the important role of multistakeholder mechanisms and expertise.



CASE STUDY 3

Systems/outcomes-based approach – New Zealand Code of Practice



General information

1. High-level description of the case study

The Aotearoa New Zealand Code of Practice for Online Safety and Harms¹¹ is a voluntary industry code that provides a self-regulatory framework aimed at improving users' online safety and minimizing harmful content online with a focus on organizations providing online services to people in Aotearoa New Zealand.

The code is intended “to provide best practices for a broad range of products and services, serving diverse and different user communities with different use cases and concerns. As such, it provides

flexibility for potential signatories to innovate and respond to online safety and harmful content concerns in a way that best matches their risk profiles”.

The code was developed between April 2021 and March 2022 by Netsafe – an independent, non-profit online safety organization, that provides online safety support, expertise and education to people in Aotearoa New Zealand – in collaboration with industry and consultation with Māori advisers, government, civil society and the public. The code was initially drafted with the involvement of major digital platforms, including Meta (Facebook and Instagram), Google (YouTube), TikTok, Twitch and Twitter, who are current signatories.



2. Context and main goals of the case study

The code addressed the following **safety and harmful content themes**:

1. Child sexual exploitation and abuse
2. Bullying or harassment
3. Hate speech
4. Incitement of violence
5. Violent or graphic content
6. Misinformation
7. Disinformation

The prioritization of these themes was informed by research conducted by Netsafe, as well as consultations with a diverse range of groups and

stakeholders, including Māori cultural advisers, civil society representatives, and academic and policy experts who have expertise on the various forms of harm. Their feedback was taken into account during the development and drafting of the code. Netsafe’s research examined how certain types of content have negatively affected people’s lives in New Zealand, as well as the top trends of harmful content reported to Netsafe over the years. It corresponds to the World Economic Forum’s risk assessment framework as follows:

(0) Identify risk:

The code calls on signatories to undertake an initial analysis of their risk profiles across the content themes and focus on systems, policies and processes that enable them to “responsibly balance safety, privacy, freedom of expression and other fundamental values”. Upon signing the code, signatories are required to submit either an initial assessment of the practices they are currently undertaking for each measure, or an explanation for why certain measures are not being implemented.

“ Signatories commit to supporting programmes and initiatives that seek to educate, encourage critical thinking and raise awareness.

(1 and 2) Reduce risk and mitigate harm:

Reduce the prevalence of harmful content

online: Signatories commit to implementing policies, processes, products and/or programmes that seek to promote safety and mitigate risks that may arise from the propagation of harmful content online while respecting freedom of expression, privacy and other fundamental human rights. This might include measures to prevent known child sexual abuse material from being made available on their platforms, to protect children against predatory behaviours like online grooming, and to reduce or mitigate the risk to individuals (minors and adults) or groups from being the target of online bullying or harassment. Signatories' measures should also aim to reduce the prevalence of hate speech, content that incites violence, violent or graphic content and misinformation, as well as raising awareness of tools for users to report content that furthers online harm. In addition, signatories commit to supporting programmes and initiatives that seek to educate, encourage critical thinking and raise awareness on how to reduce or stop online bullying or harassment, the spread of online hate speech, the spread of online content that incites violence and the spread of misinformation. Finally, they commit to collaborating across the industry and with other relevant stakeholders to respond to evolving threats of child sexual exploitation, harms arising from online hate speech, content that incites violence, misinformation and disinformation.

Empower users to have more control and make informed choices:

The signatories of this agreement acknowledge that every online user has different needs, sensitivities and tolerance levels that shape their online experiences and interactions. Therefore, a single set of standards may not suffice to meet the diverse requirements of all users and safeguard their interests. To address this issue, “empowering users to make informed choices about the content they see and/or their experiences and interactions online”. Signatories will facilitate this by offering policies, procedures or products that enable users to make informed decisions about the content they view or the advertisements they encounter. Additionally, signatories will support the dissemination of accurate and reliable information on important social issues as well as providing users with tools, programmes, resources and services that enhance their online safety.

Enhance transparency of policies, processes and systems:

The signatories of this agreement pledge to be transparent about their policies, procedures, and systems related to online safety and content moderation. This transparency is essential for building trust and promoting accountability among

users. However, the signatories recognize that there may be situations where the benefits of transparency are outweighed by the potential risks to user privacy or the security of online systems. Therefore, they will balance the need for public transparency with the potential risks to users and systems. To promote transparency, the signatories will make their safety- and harm-related policies, terms of service and information on measures to reduce the spread of harmful content accessible to users. Additionally, they will publish periodic transparency reports that include key performance indicators (KPIs) or metrics demonstrating the actions they have taken based on their policies, procedures and products to reduce the spread of harmful content online. Furthermore, signatories will submit an annual compliance report to a designated code administrator outlining the measures they have implemented and the progress they have made in fulfilling their commitments under the code.

Support independent research and evaluation:

Signatories pledge to support or engage in research activities that investigate the effects of safety interventions and harmful content on society. Signatories will also participate in events that foster multistakeholder dialogue, especially with the research community, on topics related to online safety and harmful content. Furthermore, signatories will select an independent third-party organization to review the annual compliance reports submitted by signatories. This organization will evaluate the progress made by the signatories against their commitments, outcomes and measures, as well as the commitments made in their participation form. This evaluation process will ensure that signatories are held accountable for their commitments under the code and that their efforts to promote online safety and reduce the spread of harmful content are effective.

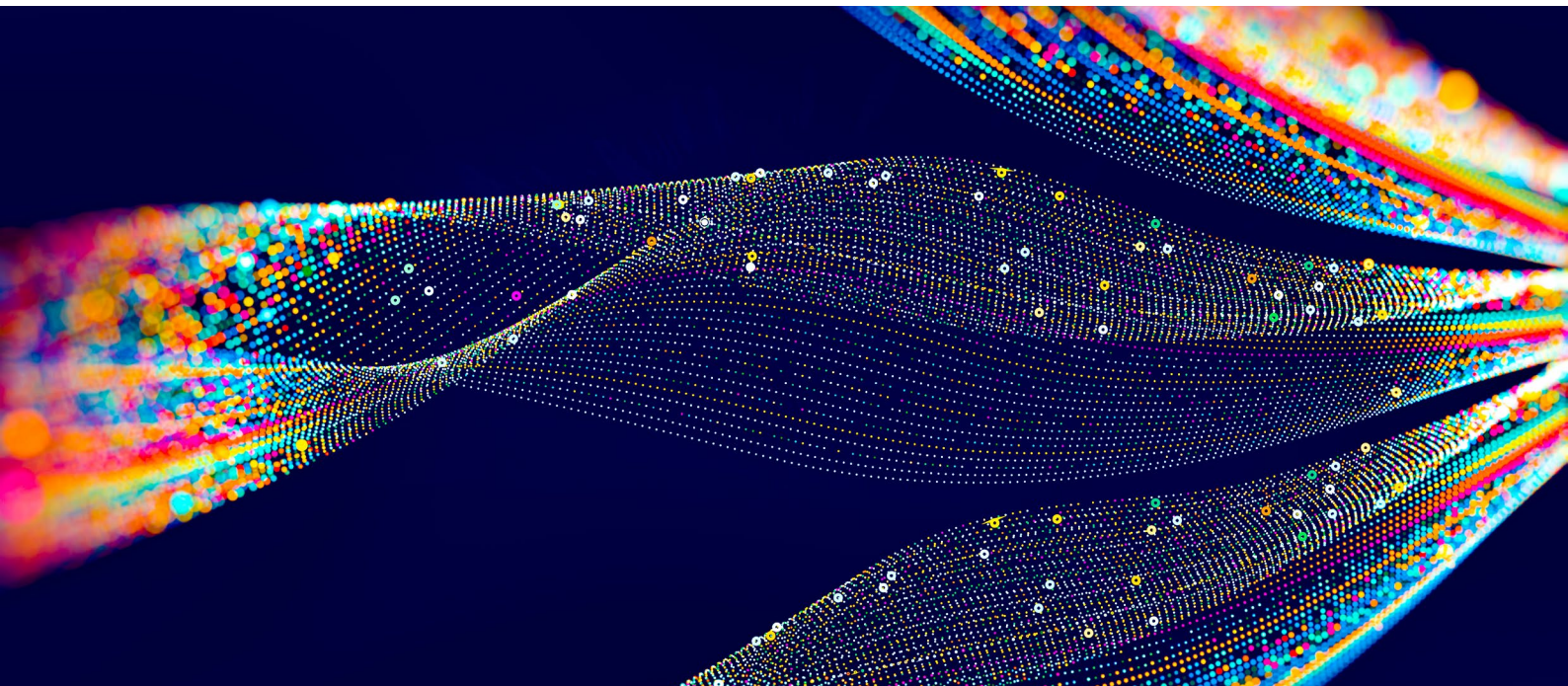
(3 and 4) Repair harm and report:

The code provides a governance framework that helps relevant stakeholders, as well as the public to hold signatories to their commitments. Per the code's own words: “Although voluntary, digital platforms that become signatories commit to being held accountable. For this purpose, the code introduces oversight powers for an administrator and a multistakeholder oversight committee. The oversight committee may recommend to the administrator the termination of a signatory's membership or the public naming of a signatory for failing to meet its commitments. In contrast, the administrator may make binding decisions. A complaints mechanism, allowing users to report on signatories' non-compliance with code commitments will also be established”.

3. Target population(s) affected by the intervention and anticipated impacts

The code was designed for organizations providing online services to people in the country and it “aims to provide best practices for a broad range of products and services, serving diverse and different user communities with different use cases and concerns”. Guiding principles include:

1. Promote safety
2. Respect freedom of expression and other fundamental human rights
3. Protect user privacy
4. Recognize the transnational or global nature of the internet
5. Broad applicability and participation
6. Systems-based best practice standards
7. Proportionality and necessity
8. Whole-of-society collaboration and cooperation.



Existing methodologies/frameworks

4. Relevant existing risk methodologies and frameworks taken into consideration

According to the drafter’s own words: “The code is an evolution of existing industry principles and standards that aims to broaden efforts, transparency and accountability for online safety and harm. It is built on existing practices in Aotearoa, New Zealand and codes of practice in other parts of the world, mainly the EU Code of Practice on Disinformation, the EU Code of Conduct on Countering Illegal Hate Speech Online, the Australian Code of Practice on Disinformation and Misinformation and the Digital Trust & Safety Partnership BPF. Most of the digital platforms involved in the code’s development are already signatories to or members of these other codes”.

The development of the code has been informed through a consultation process engaging a wide range of stakeholders across government, civil society and the public. It seeks to acknowledge and integrate local culture and principles for collaboration by incorporating Māori principles thereby creating a basis for trust, collaboration and further evolution. Those principles include:

- **Mana:** integrity and respect
- **Kauhanganuitanga:** balance
- **Mahi tahi:** working together, sharing responsibility, collaboration, cooperation and teamwork
- **Mana tangata:** showing respect, generosity and care for others.



5. Measurement frameworks and assessment metrics involved

The code of practice includes eight guiding principles, which provide a set of values to guide signatories and the administrator of the code: “The principles aim to ensure that the nature and benefits of the internet, as well as international human rights principles, best practices and standards, are taken into account”.

The eight guiding principles inform four commitments and corresponding outcomes and measures aimed at addressing concerns related to safety and the spread of harmful content online. Through their annual reports, signatories identify which of the 13 outcomes and 45 measures are relevant for their services and then report to the administrator how they are making progress towards those goals on an annual basis. Signatories are encouraged to provide relevant metrics to demonstrate progress; for example, the number of pieces of content removed for violating relevant policies or the number of people who participated in education programmes.

6. Legal or regulatory obligations and which played a role in this case study

The code does not aim to replace or address obligations related to existing laws or other voluntary regulatory frameworks. Rather, it focuses on the signatories’ systems, policies, processes, products and tools designed to mitigate the spread

of potentially harmful content. It represents an advancement of industry principles and standards, with the objective of enhancing efforts towards online safety and harm reduction, while increasing transparency and accountability for online safety.

7. Benefits and risks associated with the approach taken

The code supports cross-industry initiatives aimed at enhancing online safety. Some of the key benefits of the approach taken of participating in a cross-industry development of such a code includes:

- The code takes a systems- and outcomes-based approach towards online safety and content moderation. It facilitates accountability through transparency of policies, processes, systems and outcomes. Rather than implementing interventions that may quickly become outdated or irrelevant in the rapidly changing digital ecosystem, platforms should focus on establishing adaptable measures.
- The code applies broadly and provides flexibility to all signatories (large and small, offering a variety of products and services) to respond and comply in a way that best matches their risk profiles.
- The code facilitates accountability through transparency reporting that helps certify if a signatory exceeds, meets or falls short of code commitments.

“ The code facilitates accountability through transparency reporting that helps certify if a signatory exceeds, meets or falls short of code commitments.

- The code supports the broader policy and legislative framework by focusing on the architecture of systems, policies and processes that complements existing laws, as well as potentially bridging gaps in the legal system where the law is still in development.
- The code facilitates multistakeholder dialogue and collaboration by formalizing regular touchpoints and information exchanges between government, industry, civil society and other relevant stakeholders via its governance framework.

The risk of developing a unique code for each country lies in foregoing consistency with principles central to the functioning of a free and open internet, such as multistakeholder dialogue and collaboration or universal and open access to content from around the world, for local applicability. The approach used to mitigate the risks associated with developing a code of this nature was to engage wide arrays of communities and stakeholders through the consultation process. The active involvement of the industry ensures that technical requirements and obligations under the code are appropriate, feasible and suitable for the local market while also aligning with global industry standards.

Implementation

8. Changes from the current state or practice that resulted from the risk assessment undertaken

The launch of the New Zealand Code of Practice led to the inclusion in transparency reports of online service providers of a New Zealand-specific focus, providing visibility to the community on policies enforcement, data requests handling and intellectual property protection.

9. Investment required in terms of resources and timeline for implementation

Human resources were required both to participate in the development of the code and to ensure implementation and compliance with commitments. This involved engagement by Netsafe and other contributors for drafting and consultation periods. There is also an associated financial cost with the maintenance of the administrator and the complaints facility.

For signatories, there are substantial human resources, as well as technological resources, involved in the development of products and safety mitigations to address harms as well as to produce data for the publication of transparency reports.





CASE STUDY 4

Safety by design – The Australian eSafety Commissioner’s Safety by Design start-up assessment tool



General information

1. High-level description of the case study

This case study focuses on a fictitious social media service with a target user base of 13-18-year-olds. In this case study the Australian eSafety Commissioner's assessment tool was applied for start-up companies¹² to measure the level of user safety and to be informed about safety gaps that the platform should mitigate. The below responses represent the state of safety by design of the platform, as reflected by the assessment tool. Each answer has a label that corresponds to the risk assessment stage, referred to in question two.

2. Context and main goals of the case study

The eSafety Commissioner offers two comprehensive interactive and dynamic assessment tools that guide and support the industry to enhance online safety practices. For each assessment tool, users are provided with an educative module on online harms and taken through a series of question and response options, culminating in a tailored end report. The report acts as a safety health check and a learning resource, outlining areas to bolster safety considerations and stimulate further innovation. The report is downloadable with links and resources to refer to in future. Practical templates, example workflows, case studies and videos from leading tech experts are interspersed throughout. These provide a broad range of practical educative materials for a variety of audiences.

“ Users are provided with an educative module on online harms and taken through a series of question and response options, culminating in a tailored end report.

This case study used the tool for start-ups, which aims to provide foundational support and guidance for online platforms to enhance online safety practices by employing a wide range of approaches applicable to all touchpoints within an organization, irrespective of size and structure.

The assessment tool provides a well-rounded framework that addresses risks along the “risk life cycle”, as detailed below.

0 – Risks are identified through a series of Q&As.

1 – Practical tools, e.g. a business model canvas and risk report, are provided to support the reduction of risk.

2 – Options to mitigate harms are presented through, for example, case studies.

3 – Repairing of harm is addressed by guidance on, for example, internal operational guidelines.

4 – Reporting is covered in guidance on, for example, reporting mechanisms.

3. Target population(s) affected by the intervention and anticipated impacts on them

Everyone (including non-users). For example, users, employees and contractors internationally.

Existing methodologies/frameworks

4. Relevant existing risk methodologies and frameworks taken into consideration

A number of models and research projects were used to categorize the types of risks and harms addressed by the eSafety Commissioner's Safety by Design principles, assessment tools and guidance materials. Much of the research that focuses on “online risks” has centred on children and young people, with a number of classification models and theories emerging that have been captured in the Safety by Design resources.

A broad array of alliances, coalitions, frameworks, guidance, codes of practice and principles focused

on online safety have been developed globally since the early 2000s. The main objective of these initiatives is to protect young people online and help parents and guardians protect their children online. The Safety by Design principles and tools drew on this objective – balanced against privacy and security.

Ethical and human rights standards and concepts were also used to underpin and guide the development of the principles and guidance materials, including assessment tools. This is in line with the work being progressed by the Australian Human Rights Commission, as outlined in their issues paper, published in July 2018. All the Safety by Design principles and resources are complementary, rather than mutually exclusive.

While not an exhaustive list, some of the frameworks and guidance includes:

- Privacy by Design/Security by Design¹³
- EU Kids Online: Final Report¹⁴ – the 4C’s
- *Luxemburg Guidelines*¹⁵
- European Parliament, *Research for CULT – Child safety online: definition of the problem*¹⁶
- Misha Teimouri et al., “A Model of Online Protection to Reduce Children’s Risk Exposure: Empirical Evidence from Asia”¹⁷
- UNICEF, *Children in a Digital World: The State of the World’s Children 2017*¹⁸
- Ofcom, *Addressing harmful online content: A perspective from broadcasting and on-demand standards regulation*, 2018
- The Berkman Center for Internet & Society at Harvard University, *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.
- Global Kids Online and the DQ Institute research and impact reports for an overview of research on risks and harms faced by children and young people globally.¹⁹
- *ReCharge: Women’s Technology Safety, Legal Resources, Research & Training*²⁰
- Europol, *Internet Organised Crime Threat Assessment*²¹
- Child Dignity Alliance, *Child Dignity Alliance: Technology Working Group Report*, n.d.
- Youth Vision Statement – The views and lived experiences of young people, through Safety by Design workshop²²



5. Measurement frameworks and assessment metrics involved

The first step was to run a [questionnaire](#) that assesses the safety mechanisms of online platforms and gives a clear understanding of where safety risks exist.

Out of the 21 assessment tool questions:

- Ten of the questions touch on the initial stage of **identifying risks and risk factors (stage 0)**, such as who the user base is (e.g. minors), whether the company has community standards in place and whether it has a dedicated team to deal with safety concerns. Answering “no” to these questions puts the company at a higher level of risk, as it doesn’t have the most basic safety mechanisms in place.
- Eight of the questions deal with **reducing risks (stage 1)**, such as having safety reviews as part of the product design, having content moderation processes in place and implementing tools to assure age-appropriate access to content.
- Four of the questions deal with **mitigating harm (stage 2)**, such as having user reporting mechanisms in place, and having policy violation enforcement mechanisms.
- Three of the questions deal with **repairing harm (stage 3)**, such as having support services for employees who review harmful content and having moderation practices to manage user behaviour.
- Two of the questions deal with **reporting harm (stage 4)**, such as having a transparency report in place, and having a reporting mechanism to law enforcement.

“ By answering the questionnaire, an online platform can immediately see where it has gaps that may hinder its efforts in safety and user trust.

In addition to the risk assessment set of questions, the Safety by Design toolkit provides additional frameworks to help with the following:

Reducing risk: The **business model canvas** lays out a set of additional questions that a company should address to flesh out its position towards user safety.

Identifying risk: The **risk threshold** exercise helps companies understand the risk factors of their users and their interactions, based on personal attributes and behavioural attributes. The **staff training guide** gives a framework for training sessions to equip employees with a deep understanding of user safety, from regulatory requirements to content moderation procedures.

Additional resources are also provided within the toolkit, including:

- A **content moderation workflow**, which lays out all the stages of the process, from mapping regulatory requirements to defining policies to enforcing methodologies. This helps with risk assessment from a bird's-eye view.
- **Start-up questions and answers**, which provide an additional framework to help companies build standards around content moderation, policy development and reporting mechanisms.

6. Legal or regulatory obligations that played a role in this case study

A few regulatory frameworks inspired the eSafety Commissioner's Safety by Design assessment tool. All of them focus on general responsibilities of online platforms to secure the safety and well-being of their users, including children. Although indirectly relevant, topic-specific regulations that focus on specific harms like misinformation or terrorism were excluded. With that in mind, the following laws or proposed legislations were examined: the Digital Services Act in the EU, UK's Online Safety Bill, Singapore's Online Safety Bill, Australia's Online Safety Act and Ireland's Online Safety and Media Regulation Act.

7. Benefits and risks associated with the approach taken

Benefits of safety by design include:

- Taking a human-centric approach that places the safety and rights of users at its core, while also taking into account their needs and expectations.
- Accounting for the needs of other participants in the technology ecosystem through multistakeholder consultations with NGOs, advocates, parents and young people.
- Outlining realistic, actionable and achievable measures that providers of all sizes and stages of maturity can use to safeguard users from online risks and harms.
- Providing a structured framework to embed user safety into systems and processes and mitigate risks ahead of time.
- Incorporating a significant amount of multistakeholder input as a result of in-depth consultation with large technology companies and early stage or start-up companies.
- Promoting the technology industry's strengths in innovation, encouraging new thinking and investment that supports product development that prioritizes online safety.

By answering the questionnaire, an online platform can immediately see where it has gaps that may hinder its efforts in safety and user trust. The business canvas sets the right questions to help keep the company focused on high-priority safety issues.

There is a gap in the lack of a structured framework for policy development, which can be difficult for start-ups and small companies who don't have expertise in this area. The risk threshold document mentions contact and content risks, detection and behavioural activities but, even when combined with the safety by design typology of harms and guidance and video content on policies, it may not be enough to guide companies to create a policy based on their values and mission.



Implementation

8. Changes from the current state or practice that resulted from the risk assessment undertaken

The questionnaire in this case study would help the fictitious company reveal its most pressing safety risks and inform prioritization of mitigation actions. For example, where the questionnaire showed that the platform lacked a transparency report, the risk was framed as of a lower urgency than the risk revealed in the question about having policies in place (or not). Having a policy or community guidelines in place is a fundamental part of building a safety mechanism, as it defines the core values the platform wishes to uphold. Receiving a clear output of all the safety risks ranked by urgency, based on the Safety by Design questionnaire, enables a company to create a structured approach towards building a safer platform, even if it lacks knowledge or experience in trust and safety.

9. Investment required in terms of resources and timeline for implementation

The initial assessment is likely to involve 1-3 people and take 2-3 hours. Implementation of findings

requires resources from across the business and may take 3-6 months. It might also require external expertise or consultancy, for example for start-ups or smaller companies that don't have requisite expertise in-house. The framework should be subject to continual review and improvement.

10. Other outcomes

The framework can help prevent public relations crises, protect the platform's reputation and increase trust in key parts of the intended audience (e.g. children, but, perhaps more importantly, parents/carers and schools).

The assessment tool shares best practices and innovations other companies have used in terms of addressing a broad range of online safety problems. It also takes a "whole of organization" approach reinforcing that leadership begins at the top and a culture of safety must be embedded throughout the organizational structure. Specific resources, such as the Business Model Canvas, lay out how this can be achieved.

These resources have been drawn on to inform the Australia's national Digital Technology Curriculum and are also used in interdisciplinary subjects by various universities.



CASE STUDY 5

Child safety – gaming, immersive worlds and the metaverse



General information

1. High-level description of the case study

This case study focuses on the dynamic and immersive elements of metaverse/gaming in comparison to “traditional” social media and gaming experiences, with a focus on child-related risks on the platform. It covers key aspects of the end-to-end user experience, including user registration, payment methods and

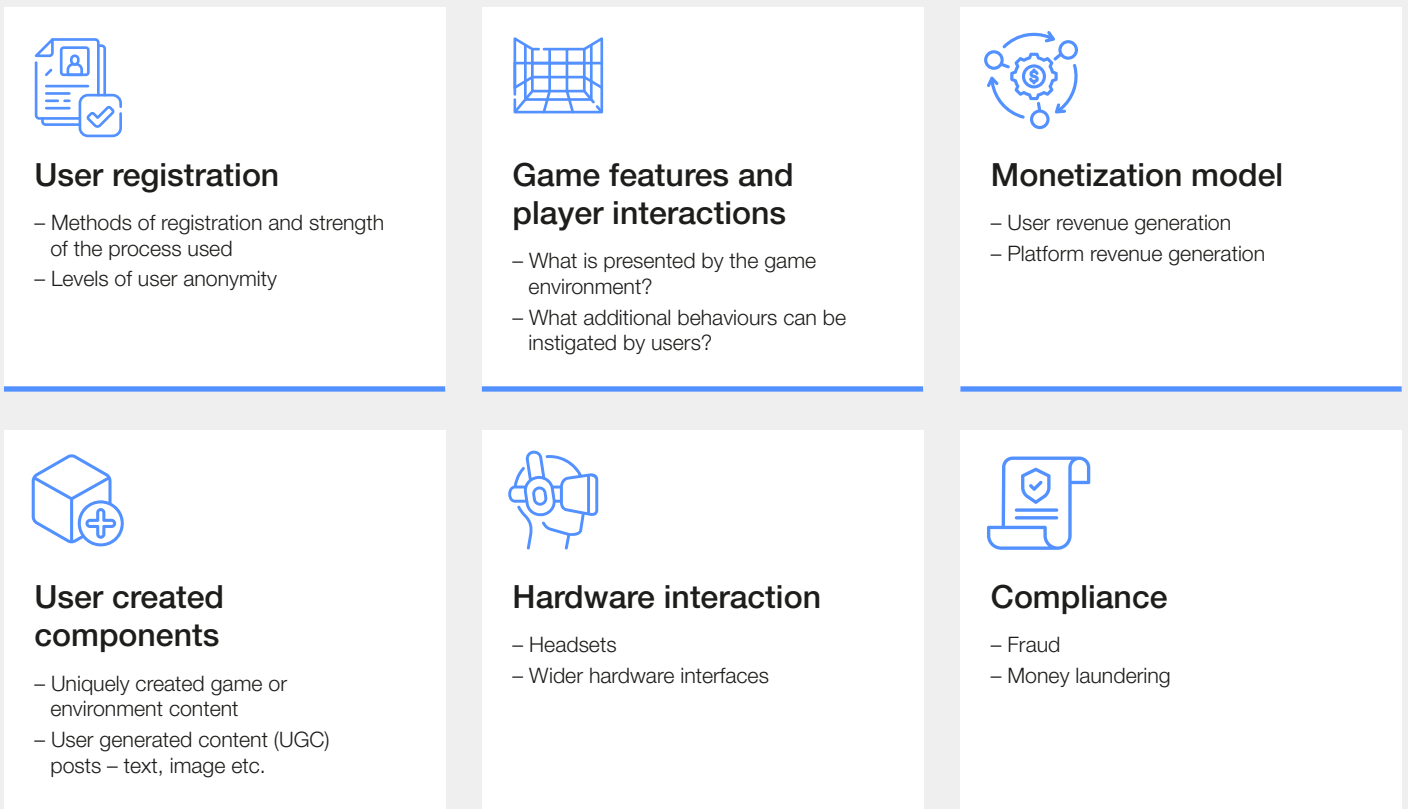
player interactions. It follows a typical user journey from user registration through to avatar definition and wider creative features, but also covers several wider compliance factors such as fraud detection.

Overarching structure

The following section is a subset of the Crisp, a Kroll Business Gaming Risk Framework. It should be read in conjunction with the eSafety Commissioners framework.

FIGURE 6

Crisp, a Kroll Business: Risk Assessment Framework, gaming and metaverse – high-level



An internationally recognized definition of a child is provided by UNICEF in the United Nations Convention on the Rights of the Child, stating that “a child is defined as every human being below the age of eighteen years, unless under the law applicable to the child, majority

is attained earlier”. This represents a portion of the assessment criteria and focuses on the key considerations of gaming and metaverse environments, critically those that are unique when compared to traditional social media environments.

TABLE 1 | Key areas assessed/covered

Low      High












Section reference	Title	Rationale	Risk assessment outputs (conducted for a gaming client)	Risk score
1	User registration	Methods for user registration and respective strength of the associated registration methods	<ul style="list-style-type: none"> – Weak registration (e.g. not requiring any form of account verification such as through email) – Ability to create multiple accounts from a single device 	
2	Payment methods	For platforms that have a payment mechanism, the type and strength of registration	<ul style="list-style-type: none"> – Credit card payment and crypto payments aligned to a single account 	
3	Commercial model	Free to play or other models that provide different payment models	<ul style="list-style-type: none"> – Free to play – the ability to pay for advancement within the game from initial registration and throughout the game (the ability to pay for progress could make this vulnerable to bad actor behaviours) 	
4	Online vs offline	Gameplay completely offline, hybrid or fully online	<ul style="list-style-type: none"> – Game is fully online – there is no offline version available – always connected model could potentially be higher risk as it is more dynamic vs static so risks may evolve over time 	
5	Platform age assessment (PEGI)	Formal age published for the game against PEGI or similar depending on the region. Digital age of consent may vary by country	<ul style="list-style-type: none"> – PEGI 12 – although there is significant user-generated content (UGC) within the game. 	
6	Intermingling in-game	Intermingling in the game – adults and children co-playing or interacting in the game, intentional or otherwise	<ul style="list-style-type: none"> – Yes, as it is PEGI 12 – therefore by design – adults can register and engage in the same spaces as minors – this is evident in wider game context 	
7	Player interactions	Policies explicit or implicit around player interactions and player discovery	<ul style="list-style-type: none"> – Players can interact with any player in their self-selected environment (i.e. UK/US regional users – normally managed through a combination of geo-fencing and language selection) – Individual to group (public room) – Options for private invite only rooms – Private audio streams 	
8	Status generation	Consideration of status generation within immediate gameplay or the wider game environment.	<ul style="list-style-type: none"> – Users gains status through the completion of challenges – this results in the award of player enhancements and items – Wider player enhancements are purchasable within the game – there is no need to spend time building up a significant proportion of player enhancements 	
9	Ranking systems	Public rankings of players – visibility of this across the game – by tribe/guild, region/language or pan environment	<ul style="list-style-type: none"> – Game has a public rankings system – where peers/players are compared against each other 	
10	Currency/wealth creation (virtual or real world)	In-game wealth generation model characteristics and methods to accelerate wealth creation or ownership	<ul style="list-style-type: none"> – There is the ability to convert real-world currency into in-game currency and vice versa – Items can be purchased and then exchanged within public, private groups and direct messages (DMs) 	
11	Item collection/management	Item policies and in-game mechanics in relation to item management, including creation and enhancement	<ul style="list-style-type: none"> – There are scarce or rare items within the game that have monetary value and status within the game 	
12	Exchange or conversion of value	How is value defined and how is it exchanged or transferred?	<ul style="list-style-type: none"> – Gifting is possible within the game – between two users – although both must confirm to initiate and complete the gifting transaction 	

TABLE 1 | Key areas assessed/covered (continued)

Low ● ● ● ● ● High

Section reference	Title	Rationale	Risk assessment outputs (conducted for a gaming client)	Risk score
13	Collaboration for common goals	Common themes or game mechanics for player discovery of new players across the environment	– There are options to collaborate with strangers on short mission-based common goals – this enables “on-mission” private room chat	●
14	Marketplaces	Linked to the economic model of the environment – are there in-game or on-platform marketplaces?	– In-game marketplaces – enable player enhancements, items and map/environmental elements to be acquired or spent – This can be equipped and used in-game: these can be traded and exchanged in-game “between players” – usually for similarly priced items – Users can build or modify items of selling in the marketplaces – for certain classes of items users can import their own images or skins	●
15	Content creation	What are the policies and scope for content creation and “zero-day” content	– Users can create and edit avatars from a from a pre-defined set of building blocks – These designs can be saved and traded for in-game currency – Items can be created through some basic primary shapes – these can be “skinned” with user imagery – basic assessment of uploaded skins (2D vs 3D assessment problem)	●
16	Game mechanics	Any wider or unique game mechanics, which could create additional risks	– Users can purchase (with real world currency) an additional environment editor. Within that editor they are equipped with basic building blocks for scene creation. They can also create objects for interaction in the game – effectively building their own missions	●
17	Proactive identification of harms (child)	What level of proactive identification is undertaken by the platform	– Early warning risk detection provider providing proactive assessment of emergent threats	●
18	Hardware interfaces	How is the environment presented – i.e. what type of end-point user device and what additional child risks may this provide?	– Multiplatform – desktop and tablet – no hardware risks assessed in current configuration (basic camera risk on tablet and off-platforming to a less controlled environment)	●
19	Criminal behaviour/fraud prevention	Does the platform already have a fraud protection or anti-money laundering (AML) function	– The game has AML and fraud detection teams – linked to their credit card payment solutions and to their crypto payments approach.	●

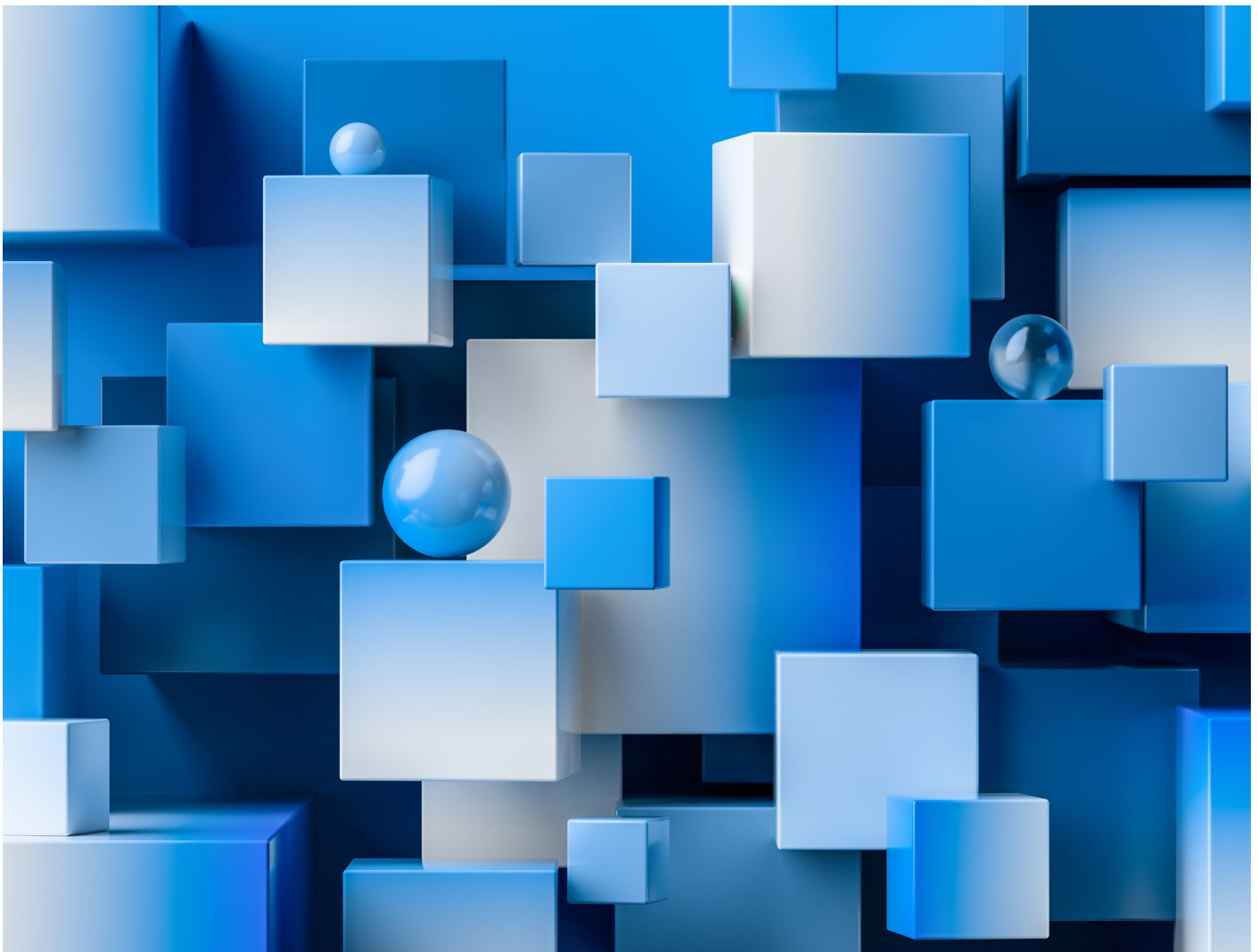
2. Context and main goals of the case study

The case study is based on a real yet confidential client assessment performed in April 2019 by a Crisp risk consultancy team. This gaming platform is:

1. A 3D world, with the ability to edit character/avatar, items, environment, game economics and overall game mechanics
2. Rated PEGI 12 across a range of app stores and online marketplaces as per the International Age Rating Coalition’s (IARC) classification system

3. Should be considered an intermingled environment – where all ages of users can interact with all users
4. The game has several economic models within the gameplay – including in-game markets and external markets.

This case study is focused on the “identify risk” phase of the risk assessment framework, focusing specifically on risks related to child safety, particularly CSAM.



3. Target population(s) affected by the intervention and anticipated impacts

The risk assessment conducted focused on identifying and mitigating risks to children, with interventions aimed at improving the game design and associated game mechanics.

Existing methodologies/frameworks

4. Relevant existing risk methodologies and frameworks taken into consideration

- Australia's eSafety Commissioner – enterprise assessment criteria.²³
- Crisp, A Kroll Business – child safety risk assessment framework (partial framework is provided as part of this case study): key focus on non-traditional social media vectors or features of gaming. The framework covers both externally observable aspects (e.g. user/player features) and internal governance and response elements, unique to gaming such as criminal behaviour or fraud prevention.

- PEGI (12)/IARC assessment of the game.

5. Benefits and risks associated with the approach taken.

The approach taken provided a clear assessment of the baseline for the platform, allowing for the creation of a heatmap, the identification of attack surfaces and the creation of a framework that allows for prioritized operational and strategic interventions.

Implementation

6. Changes from the current state or practice that resulted from the risk assessment undertaken

The overall assessment provided a first set of focus areas for triaging, including supporting the development of a business case to invest in the set-up of a trust and safety team. This included policy development and tooling for the enforcement team.

The assessment then supported the creation of a heatmap to identify areas of risk across the platform. Based on this exercise, a set of targeted interventions to reduce risk were identified, including revisions of policies across:

1. Registration methods: strengthened to reduce or sandbox unregistered users
2. Multi-accounts per device: reduction in number of accounts per device
3. Player interaction: reducing maximum number of stranger invites and additional social graph analysis, updated policy to filter minors' personally identifiable information (PII) from chats, including location, online handle or telephone details
4. Private space/chat controls: reduction on the pace at which the new players can engage unknown or strangers
5. Gameplay time: developed a connection between the time played and availability of features and volumes of requests and interactions possible ("trust laddering" model)

6. Avatars and item skinning: introduced constraints such as the request of a JPEG upload

7. Gifting: updated game mechanics and scripts/missions to limit gifting to only necessary circumstances within the game, e.g. not permitted to minors and not from strangers.

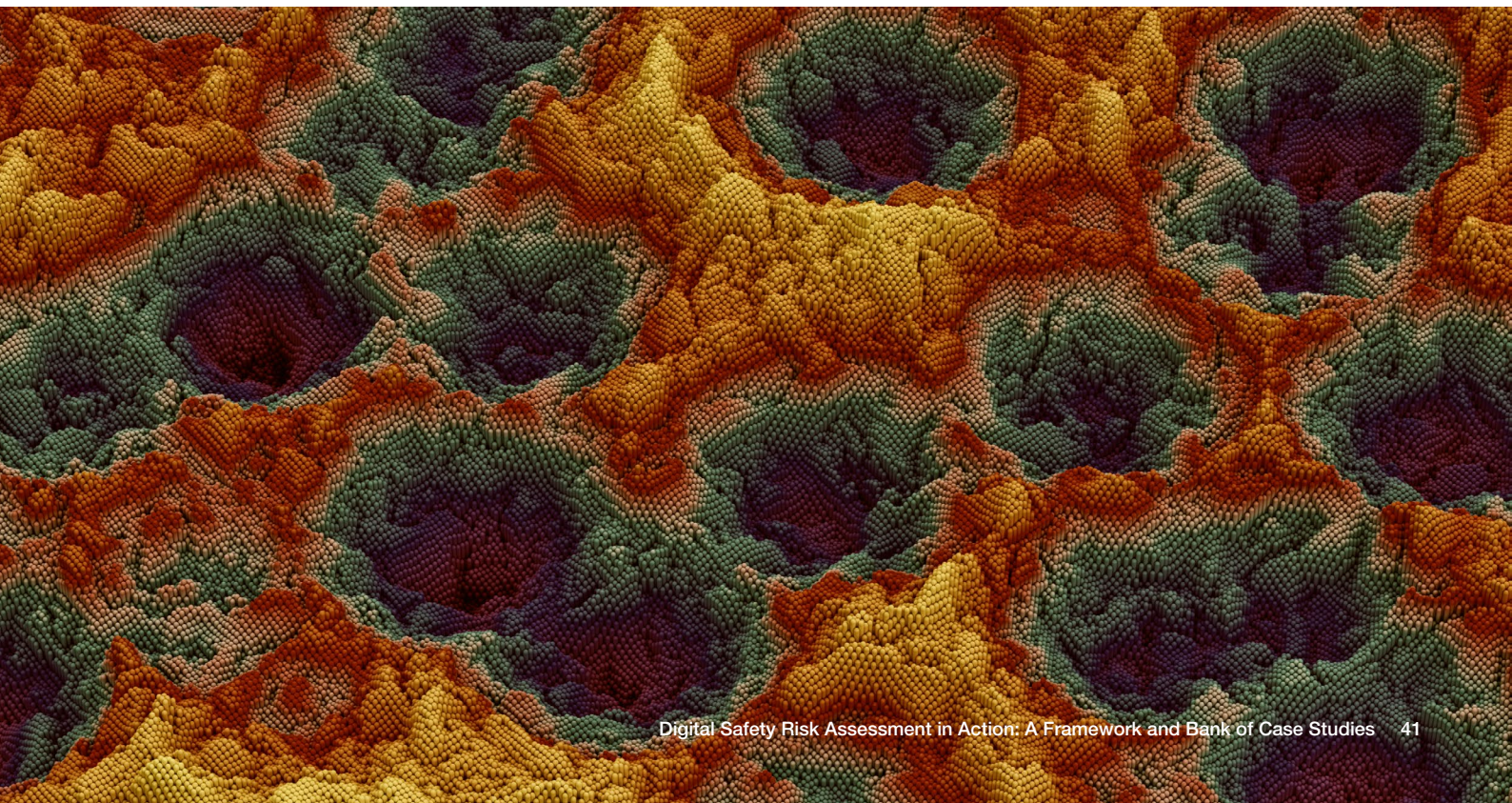
Several adult sexual content policies were also enacted to improve immediate child safety, for example addressing asserted adults discussing sexual roleplay.

7. Investment required in terms of resources and timeline for implementation

One to two full-time employees were required for initial and ongoing assessment and policy development. The findings should be reviewed on a six-monthly basis or as required based on feature releases, community feedback or proactive identification of issues.

8. Other outcomes

The assessment improved the company's understanding of risks to children on their service and the impact of changes to game dynamics in relation to trust and safety considerations. It has driven a better understanding of building in safety by design as part of the game's mechanics and dynamics and of the wider relationship across risks and harms, including how child safety risks differ from considerations around adults.

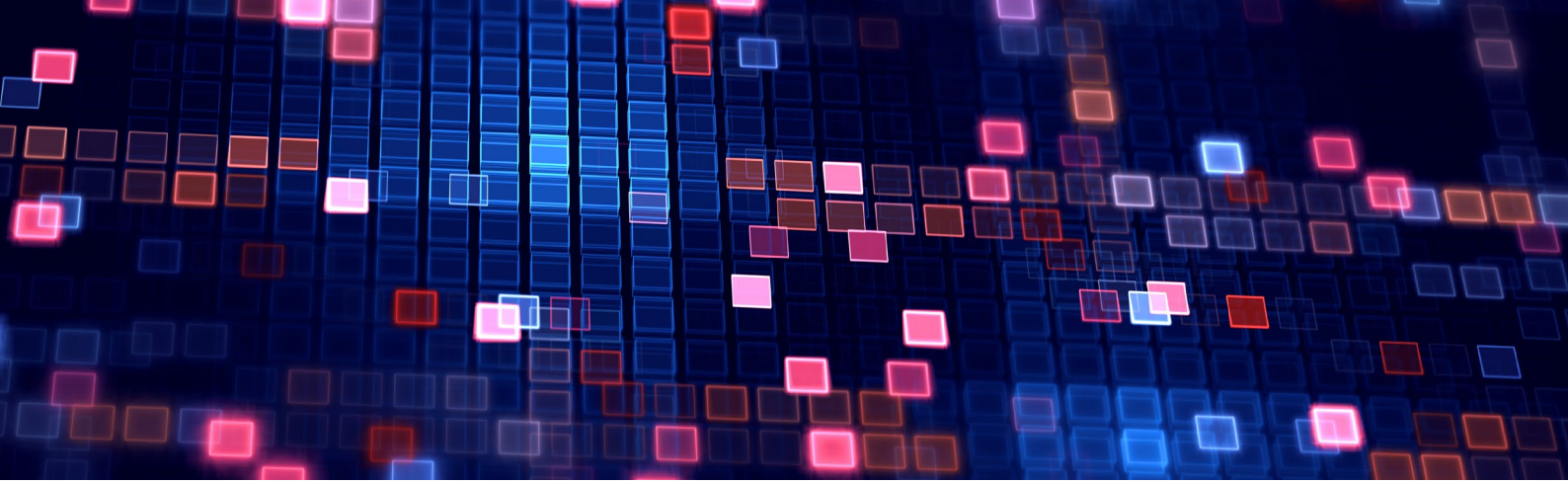




CASE STUDY 6

Algorithms – AI impact assessment tool





“ The relevance of search results may differ between each user class; thus, the nature and extent of any intervention may impact individual users differently.

General information

1. High-level description of the case study

This case examines the impact assessment process of developing new automated features for a search engine to combat the spread of undesirable content. For the purposes of the case study, these features include automatic detection of the content, automated demotion of the content in search results, and controls for humans to initiate or manage automated interventions. However, the methods used for this case study are applicable beyond these theoretical features and on platforms and services other than an extensive online search engine.

The impact assessment was performed using Microsoft’s responsible AI methodology. Resources available from [Microsoft](#) include a playbook (*Responsible AI Standard*), a guide and an impact assessment template.

2. Context and main goals of the case study

This case study discusses the implementation of automated features in a search engine. These features aim to both identify material that may harm users and reduce such risks and mitigate harm by reducing the visibility of undesired content in search results. This is particularly important in situations where users search for information in areas where there is little existing content (so called data voids) or during unexpected events that generate a lot of time-sensitive inappropriate content, such as live-streamed terrorist activity.

Context:

- Given the vast amount of information on the web, users often need a way to quickly find credible information that is relevant to their needs; web searches fulfil this goal.

Search engines build “statistical models based on previous patterns found in training data” to “quickly identify and prioritize content that most likely matches the desired goals of a searcher”.²⁴ They do this by drawing on available information such as URLs, site content, links, images and videos and derive the relevance of the results from knowledge, if any, about the end user (location, search history etc.). Existing systems for categorizing static content as adult-only existed place for a long time.

- Data voids occur when searching for terms where the available relevant data is limited, non-existent or deeply problematic. Most of these searches are rare, but when people search for these terms, search engines tend to return results that may not give the user what they want because of limited data and/or lessons learned through previous searches. Actors may deliberately exploit data voids for specific purposes.
- Search engines exist in a complex and unpredictable environment or may be subject to drifts in input distributions over time. Language and communication norms change rapidly and the many types of inputs may significantly vary in quality.

3. Target population(s) affected by the intervention and anticipated impacts

All service users are potentially affected by the intervention of the features. End users may operate search engines with known user accounts, pseudonymized accounts or anonymous accounts. The relevance of search results may differ between each user class; thus, the nature and extent of any intervention may impact individual users differently. The differing impacts on the various end user classes were considered when performing impact assessment but were insignificant.

Existing methodologies/frameworks

“ AI training materials are not equally available for all cultures and languages. These factors may impact the continuous improvement of the system.

4. Relevant existing risk methodologies and frameworks taken into consideration

The methodology is substantially the same as that used for previous versions of the service.

5. Measurement frameworks and assessment metrics involved

- The system is continuously improved to detect better and manage data voids and prevent them from being gamed by malicious actors, with the following challenge: both AI-driven and human-curated processes will depend on resources specific to region and language. Humans can't effectively curate in unfamiliar cultures and languages and recruiting and training such humans is slow and expensive. Likewise, AI training materials are not equally available for all cultures and languages. These factors may impact the continuous improvement of the system.
- In some cases, the content being scrutinized is part of a sudden, time-sensitive event, such as a live-streamed act of violence. These circumstances will exacerbate the conditions above.
- Effectiveness metrics would be developed to address both the continuous improvement aspects of the case study and the online search service's transparency/reporting goals, but details were not shared.

6. Legal or regulatory obligations that played a role in this case study

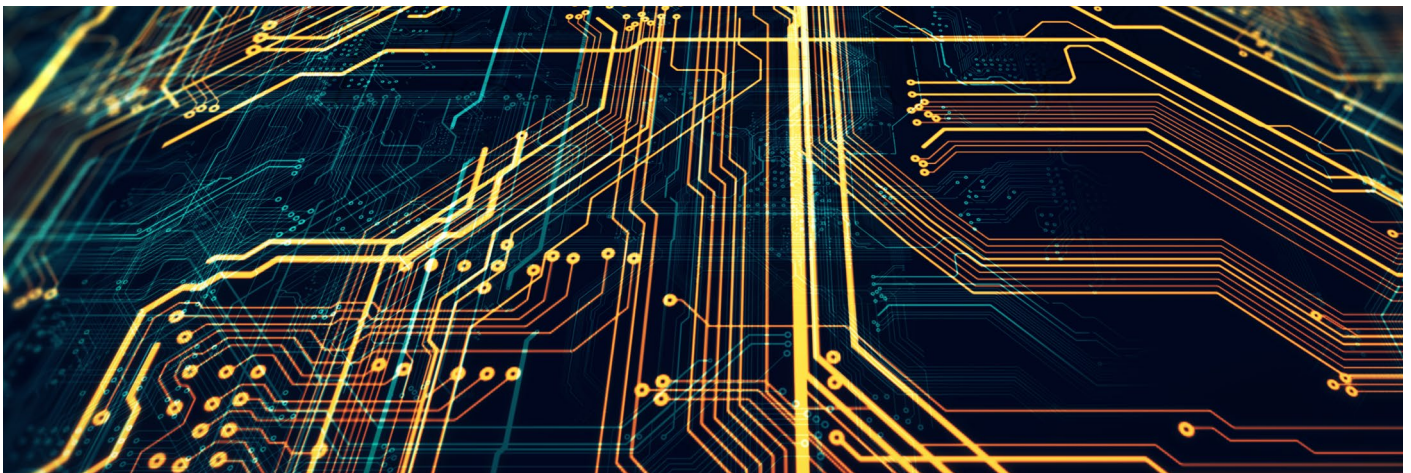
Regulatory obligations are quickly developing regarding online content in many regions. There is wide variation regarding how content is categorized and which actions service providers must take related to various categories of content. Under the EU Digital Services Act, substantial search engines must perform systemic risk assessments and audits. Alone or as part of industry partnerships (such as the DTSP), service providers are voluntarily developing risk-based best practices to align their content risk management efforts with the evolving regulatory landscape.

This changing landscape exists alongside the security and privacy environments developed in recent years. This case study focuses on impact assessment for responsible AI, but threat modelling and privacy impact assessments are also required.

7. Benefits and risks associated with the approach taken

The case study used the Microsoft Responsible AI methodology. The benefit of this approach is that it is well-aligned to the development of international standards, those developed by the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC), and was developed alongside ISO/IEC 42001 Artificial Intelligence Management System (AIMS). Multiple international standards related to AIMS are expected in risk management, governance and certification, making this a robust approach for future planning.

The methodology based on international standards is easier to develop and more resistant to fragmentation in different regions. Some regulation is influenced by existing or developing standards; some even reference standards directly.



Implementation

8. Changes from the current state or practice that resulted from the risk assessment undertaken

In current practice, a search engine regularly monitors for violations of its webmaster policy, including attempts to manipulate the search algorithm through prohibited practices such as cloaking, link spamming, keyword stuffing and phishing. A search engine provider dedicates meaningful resources to maintaining the platform's integrity, promoting high authority and relevant results, and reducing spam (including spam aimed at distributing low authority information and manipulative content). This uses a combination of human intervention and AI-driven analysis to regularly review, detect and address spam tactics occurring on search. When websites deploying manipulative techniques or engaging in spam tactics are detected, those websites may incur ranking penalties or be removed from the search index altogether.

In addition to current practice, the engine implements a defensive search and data void mitigation capability. The search algorithm endeavours to prioritize relevance, quality and credibility. Whenever it identifies a threat that undermines the efficacy of its algorithms, it employs "defensive" search strategies and interventions to counteract threats per its trustworthy search principles to protect search users from: 1) being misled by untrustworthy search results, and/or 2) inadvertently being exposed to unexpected harmful or offensive content. Defensive search interventions may include algorithmic interventions (such as quality and credibility boosts or demotions of a website), restricting autosuggest or related search terms to avoid directing users to problematic queries and manual interventions for individual reported issues or broader areas more prone to misinformation or disinformation (e.g. elections, pharmaceutical drugs or COVID-19).

In both the previous and new states, impact assessments of automated and manual processes are performed using threat modelling, privacy impact assessments and the AI impact assessment mentioned above.

9. Investment is required in terms of resources and timeline for implementation

Investment includes engineering and human resources. A development environment of continuous improvement is required, and resourcing must be determined appropriately.

Both AI-driven and human-curated processes will have dependencies on resources specific to region and language. Humans can't effectively curate in unfamiliar cultures and languages, and recruiting and training such humans is slow and expensive. Likewise, AI training materials are not equally available for all cultures and languages. Either of these cases could lead to the harms presented below.

10. Other outcomes

The following issues were discovered while performing the impact assessment and must be kept in mind during continuous improvement:

- In some cases, the content being scrutinized is part of a sudden, time-sensitive event, such as a live-streamed act of violence. Further, creators of problematic content may be in a race to defeat the safety features of the system so that the content will be delivered despite the system's defences.
- The impact assessment framework seeks to identify and minimize the risk of stereotyping, demeaning and erasing outputs. False positives could mask content that is not unlawful. If the system is trained to demote or block content because it's deemed problematic by a subset of users or flaggers – whether inadvertently or maliciously – then creators or consumers of the content won't be able to access it. If the creators or consumers are part of an at-risk population, such demotion may infringe on their human rights.
- Under these circumstances, the system may not perform as desired, and human intervention may be required. The system needs to evolve to continuously mitigate these changes in circumstance.

Conclusion

This report proposed a framework that can serve as a starting point for all stakeholders across different jurisdictions to establish their risk management practices, processes and governance. Although the framework is not a regulatory compliance tool, it provides a baseline common language to enable conversations between stakeholders.

The report showcased several case studies that demonstrate various approaches to risk assessments taking into consideration the impact across human rights. Some case studies focused on existing frameworks that can be used for risk assessments, while others highlighted specific use cases that stakeholders can draw inspiration from.

To complement the framework provided in this report, a typology of online harms will be produced,

providing stakeholders with a common foundational language of the wide range of online harms in scope. Additionally, a publication on risk factors, metrics and measurement will provide the useful ingredients to support stakeholders in running risk assessments, identifying risks and measuring the overall level of risk and impact of interventions. The coalition will also produce a report exploring solutions-based interventions that detail ways to reduce the risk of harm and mitigate and repair it when it occurs.

The framework presented in this report, along with the forthcoming papers, aims to support stakeholders in developing a comprehensive risk management governance framework that considers various human rights and provides solutions-based interventions. Ultimately, this will help reduce the risk of harm and create a safer online world for all.

Contributors

World Economic Forum

Minos Bantourakis

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

Cathy Li

Head, AI, Data and Metaverse; Centre for the
Fourth Industrial Revolution; Member of the
ExCom, World Economic Forum

Lead authors

Dunstan Allison-Hope

Vice-President, Business for Social Responsibility

Daniel Child

Industry Affairs and Engagement Manager,
eSafety Commissioner

Inbal Goldberger

Vice-President Trust & Safety, ActiveFence

John-Orr Hanna

Chief Intelligence Officer, Crisp, a Kroll Business

Collin Kurre

Technology Policy Principal,
UK Office of Communications (Ofcom)

Deepali Liberhan

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

Jason Pielemeier

Executive Director, Global Network Initiative

Katherine Sandell

Platform Risk Program Lead, Trust & Safety, Google

David Sullivan

Executive Director, Digital Trust & Safety Partnership

Mark Svancarek

Technical Policy & Program Manager, Microsoft

Acknowledgements

Jeffery Collins

Director Trust & Safety,
Amazon Web Services (AWS)

Ken Corish

Online Safety Director,
UK Safer Internet Centre

Justin Davis

Chief Executive Officer and Co-Founder,
Spectrum Labs

Julie Dawson

Chief Policy & Regulatory Office, Yoti

Farah Lalani

Global Vice-President Trust & Safety Policy,
Teleperformance

Huan Ting Lee

Second Director, MCI-Singapore

Benoit Loutrel

Board Member, Arcom

Carolyn Lowry

Public Policy Counsel, TikTok

Susan Ness

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

Isedua Oribhabor

Business and Human Rights Lead, Access Now

Chris Priebe

Founder and Executive Chairman, TwoHat

Akash Pugalía

Global President, Trust and Safety, Teleperformance

Jacqueline Rowe

Policy Lead, Global Partners Digital

Chris Sheehy

Research and Policy Manager,
The Global Network Initiative

Matthew Soeth

Head, Trust and Safety, Spectrum Labs

Steven Vosloo

Digital Policy Specialist, United Nations Children's Fund (UNICEF)

Sam Wallace

ISSB Technical Staff, IFRS Foundation

David Wright

Director, UK Safer Internet Centre

Charlotte Yarrow KC

Online Safety, Regulatory Legal, Apple

Production

Laurence Denmark

Creative Director, Studio Miko

Sophie Ebbage

Designer, Studio Miko

Martha Howlett

Editor, Studio Miko

Endnotes

1. Digital Trust & Safety Partnership (DTSP), <https://dtspartnership.org/>.
2. DTSP, *The Safe Framework: Tailoring a Proportionate Approach to Assessing Digital Trust & Safety*, 2021, https://dtspartnership.org/wp-content/uploads/2021/12/DTSP_Safe_Framework.pdf.
3. DTSP, *The Safe Assessments: An Inaugural Evaluation of Trust & Safety Best Practices*, 2022, <https://dtspartnership.org/dtsp-safe-assessments-report/>.
4. “The GNI Principles”, *Global Network Initiative* (GNI), n.d., <https://globalnetworkinitiative.org/gni-principles/>.
5. “Prevalence”, *Meta*, 18 November 2022, <https://transparency.fb.com/policies/improving/prevalence-metric/>.
6. “Views” in *Google Transparency Report*, Google, n.d., <https://transparencyreport.google.com/youtube-policy/views?hl=en>.
7. GNI, *GNI Assessment Toolkit*, 2021, <https://globalnetworkinitiative.org/wp-content/uploads/2021/11/AT2021.pdf>.
8. “Human Rights Everywhere All at Once”, *Business for Social Responsibility* (BSR), 8 September 2022, <https://www.bsr.org/en/blog/human-rights-everywhere-all-at-once>.
9. “Across the Stack Tool: Understanding Human Right Due Diligence (HRDD) Under and Ecosystem Lens”, *BSR*, n.d., <https://eco.globalnetworkinitiative.org/>.
10. “Company Assessments”, *GNI*, n.d., <https://globalnetworkinitiative.org/company-assessments/>.
11. Netsafe, *Aotearoa New Zealand Code of Practice for Online Safety and Harms*, 2022, <https://nztech.org.nz/wp-content/uploads/sites/8/2022/07/FINAL-NZ-Code-of-Practice-for-Online-Safety-and-Harms-25-July-2022.pdf>.
12. “Start-ups”, *eSafety Commissioner*, n.d., <https://www.esafety.gov.au/industry/safety-by-design/start-ups>.
13. Cavoukian, Ann, *Privacy by Design: The 7 Foundational Principles*, Information and Privacy Commissioner Ontario, Canada, 2011, <https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples>; “Safety by Design, Secure by Default”, *Cybersecurity & Infrastructure Security Agency*, n.d., <https://www.cisa.gov/securebydesign>.
14. Livingstone, Sonia and Leslie Haddon, *EU Kids Online: Final Report*, London School of Economics and Political Science (LSE), 2009, https://eucpn.org/sites/default/files/document/files/5_eu_kids_online_-_final_report.pdf.
15. “Luxembourg Guidelines”, *ECPAT*, n.d. <https://ecpat.org/luxembourg-guidelines/>.
16. European Parliament, *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](https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/602016/IPOL_IDA(2018)602016_EN.pdf).
17. Teimouri, Misha et al., “A Model of Online Protection to Reduce Children’s Online Risk Exposure: Empirical Evidence From Asia”, *Sexuality & Culture*, vol. 22, 2018, pp. 1205-1229, <https://link.springer.com/article/10.1007/s12119-018-9522-6>.
18. UNICEF, *The State of the World’s Children 2017*, 2017, <https://www.unicef.org/reports/state-worlds-children-2017>.
19. *Outsmart the Cyber-Pandemic: Empower Every Child with Digital Intelligence by 2020*, 2018, https://www.dqinstitute.org/2018dq_impact_report/.
20. Women’s Legal Service NSW, Domestic Violence Resource Centre Victoria and WESNET, *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>.
21. Europol, *Internet Organised Crime Threat Assessment 2018*, 2018, <https://www.europol.europa.eu/internet-organised-crime-threat-assessment-2018>.
22. eSafety Commissioner, *Safety by Design, Our Vision: Young People*, n.d., <https://www.esafety.gov.au/sites/default/files/2019-10/SBD%20-%20Vision%20for%20young%20people.pdf>.
23. “Enterprise companies”, *eSafety Commissioner*, n.d., <https://www.esafety.gov.au/industry/safety-by-design/enterprise-companies>.
24. Golebiewski, Michael and Danah Boyd, *Data Voids: Where Missing Data Can Easily Be Exploited*, *Data & Society*, 2018, https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf.



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