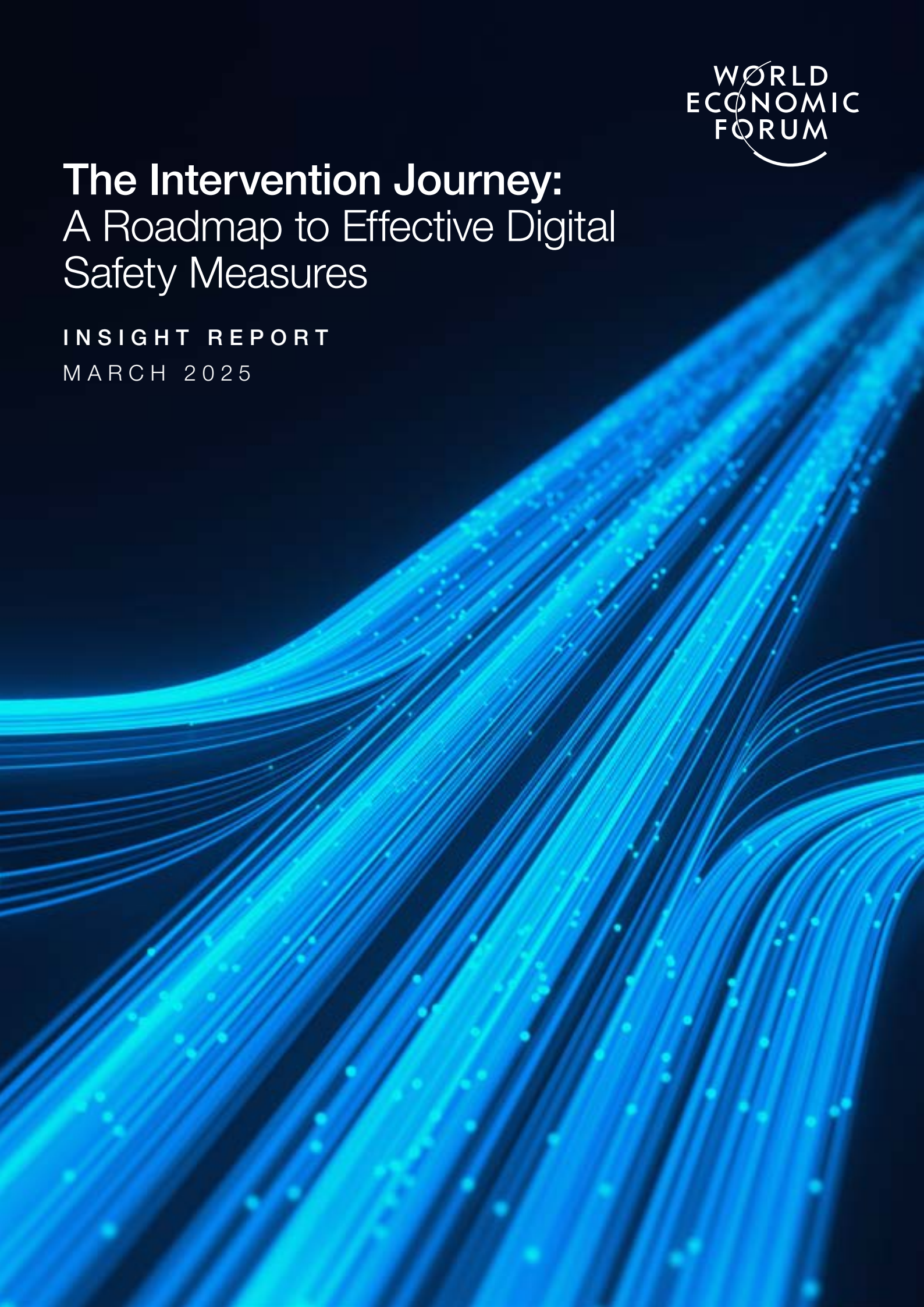
WORLD ECONOMIC FORUM

The Intervention Journey: A Roadmap to Effective Digital Safety Measures

INSIGHT REPORT

MARCH 2025



Images: Getty Images

Contents

Foreword

Agustina Callegari
Project Lead, Global Coalition for Digital Safety, World Economic Forum

Adam Hildreth
Founder, Crisp, Kroll

Daniel Dobrygowski
Head, Governance and Trust, World Economic Forum

Julie Inman Grant
eSafety Commissioner, Office of the eSafety Commissioner, Australia

In today's increasingly technology-driven world, it is essential to ensure that safety is embedded into the core of digital platforms and systems. The Safety by Design approach ensures platforms prioritize user protection, preventing harm before it occurs. Regardless of size, platforms have a responsibility to safeguard users by deploying tools like content-filtering algorithms, age verification mechanisms and accessible safety resources. As artificial intelligence (AI) becomes increasingly central to the digital landscape, prioritizing ethics and safety in its development and deployment is critical to minimizing unintended consequences.

The Global Coalition for Digital Safety recognizes the growing complexity of online harms, ranging from the production and circulation of child sexual abuse material (CSAM) to privacy violations, hate speech and disinformation. Our strategy is not to treat each challenge or risk in isolation but to develop a suite of interconnected resources.

Each tool the coalition produces complements the others, creating a global toolkit for policy-makers, industry leaders, civil society and academics.

Our Global Principles for Digital Safety and the Typology of Online Harms provide foundational guidelines and a common language for all stakeholders. The Digital Safety Risk Assessment Framework helps identify risks, while the *How to Measure Digital Safety Effectively* report outlines key categories for digital safety metrics.

This latest report, *The Intervention Journey: A Roadmap to Effective Digital Safety Measures*, provides guidance for developing effective digital safety interventions to tackle online harms. It covers the full life cycle – from identifying risks to designing, implementing and evaluating effective safety measures. Additionally, it includes real-world case studies and examples of various types of interventions. By adhering to Safety by Design principles and incorporating best practices, the report guides companies of all sizes, enabling them to manage risks associated with their digital platforms proactively.

Digital safety requires collective effort, and through collaboration with governments, industries and civil society, the coalition aims to create solutions that are globally relevant and locally adaptable.

Executive summary

The intervention journey provides a step-by-step guide for addressing online harms, ensuring safety across digital environments.

With the increased digitalization of organizations and society, along with the rising prevalence of digital harms, it is becoming ever-more urgent to combat these risks through effective digital safety interventions. This report can provide a clearer understanding of challenges and opportunities, and help organizations to identify effective strategies and cultivate collaboration and resource sharing, ultimately contributing to a safer digital environment for all users.

This report on the intervention journey builds upon the workstream of the Toolkit for Digital Safety Design Interventions, following the 2023 report *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. The journey is designed to assist organizations in navigating the complex landscape of digital safety and implementing effective and safe interventions, regardless of their size. By evaluating various intervention case studies, the report aims to offer best practices for identifying risks, selecting potential interventions and collecting and analysing feedback on those interventions.

The steps of the intervention journey are categorized as follows:

1. **Identification**: systematically recognizing and analysing factors that may threaten users, non-users and organizations in the digital environment
2. **Design**: creating targeted strategies and solutions that address specific identified risks to effectively mitigate potential harm to users, non-users and organizations
3. **Implementation**: executing the planned strategies and solutions, engaging stakeholders, providing training and monitoring the process
4. **Feedback, measurement and transparency**: collecting and analysing user and stakeholder responses to evaluate effectiveness and impact, while communicating findings

Adhering to the intervention journey ensures a structured approach to defining objectives, exploring solutions, assessing outcomes, balancing costs, integrating safety measures and evaluating effectiveness. Collaborating with partners can help organizations avoid pitfalls and solve problems more efficiently. Through open engagement with key stakeholders and sharing of best practices, solutions and technologies, partnerships can significantly enhance the impact of an intervention.

Recognizing that not all organizations, particularly small- and medium-sized enterprises (SMEs), have the capacity or expertise to effectively execute each step highlights the importance of partnerships. Improving digital safety is a cooperative effort in which collaboration benefits all stakeholders.

This fifth Global Coalition for Digital Safety publication was developed with input from its multistakeholder members, including platforms, regulators, safety providers, non-governmental organizations (NGOs), academics and international bodies. Their involvement in shaping the report underscores a shared commitment to improving the implementation of digital safety interventions across sectors and advancing efforts to make the digital space safer for all.

Introduction

In a complex digital world, proactive interventions prioritize safety while navigating regulatory demands and operational challenges.

The report guides readers through a detailed roadmap, illustrating the critical steps and considerations involved in developing and implementing effective digital safety interventions.

The digital environment is characterized by a large volume of evolving online harms, necessitating solutions that are not only adaptable and current but also proactive and forward-thinking, capable of evolving in tandem with the threats they are designed to mitigate.

In this complex landscape, a multifaceted approach is essential – no single intervention is capable of addressing the varied myriad of online harms. The nature of digital harm is such that it defies easy categorization or simple remedies. Each intervention must be carefully evaluated, weighing its costs against its benefits while also considering its effectiveness.

Digital harm lacks a definitive solution. The focus should be on mitigating risks before harms occur, while recognizing that some level of risk will always persist. Interventions in digital safety come with inherent costs that extend beyond financial expenditure, impacting areas such as time, user experience and freedom of expression. As a result, it is essential to carefully weigh these costs against the potential benefits, such as increased security, privacy and user trust, and the protection of fundamental human rights. Companies must weigh the gravity of actual harm from serious crimes such as child sexual exploitation and abuse (CSEA), since such protection is an internationally recognized fundamental right.

Interventions in digital safety are broad and encompass both technological and non-technological measures to effectively address the complexities of online threats. In all approaches, measuring and analysing digital safety interventions is crucial. It is also essential to consider how widely they can be adopted, their suitability for providers, the ease of implementation and overall effectiveness.

Interventions are increasingly shaped by evolving regulatory frameworks, including Australia's Online Safety Act 2021, the United Kingdom's Online Safety Act 2023 and the European Union's Digital Services Act. Governments and international organizations are advocating for stronger measures to protect users and vulnerable non-users, and hold platforms accountable.

However, detailed disclosures about the inner workings of digital safety tools could be increasingly exploited by malicious actors, who might use this information to better bypass security measures or manipulate systems. In this context, the challenge lies in balancing the need for transparency with the necessity of maintaining security.

Smaller companies often face challenges in implementing effective digital safety interventions due to limited technical expertise and budget constraints. This makes the sharing of information and best practices between platforms, such as through partnerships, crucial for addressing the global and interconnected nature of online harms.

This report is designed to compile a broad spectrum of digital safety interventions, providing companies with the insights needed to enhance their capabilities in identifying digital risks, preventing harm, mitigating impacts and facilitating repair in the aftermath of incidents.

The spectrum of interventions has been organized into four types, with many interventions intersecting with multiple types. Those types are: technical, behavioural, educational and policy-related interventions.

The report guides readers through a detailed roadmap, illustrating the critical steps and considerations involved in developing and implementing effective digital safety interventions.

The objective is to provide valuable insights and practical recommendations, specifically for SMEs. The sharing of best practices and use of partnerships can help them to develop and deploy robust safety measures.

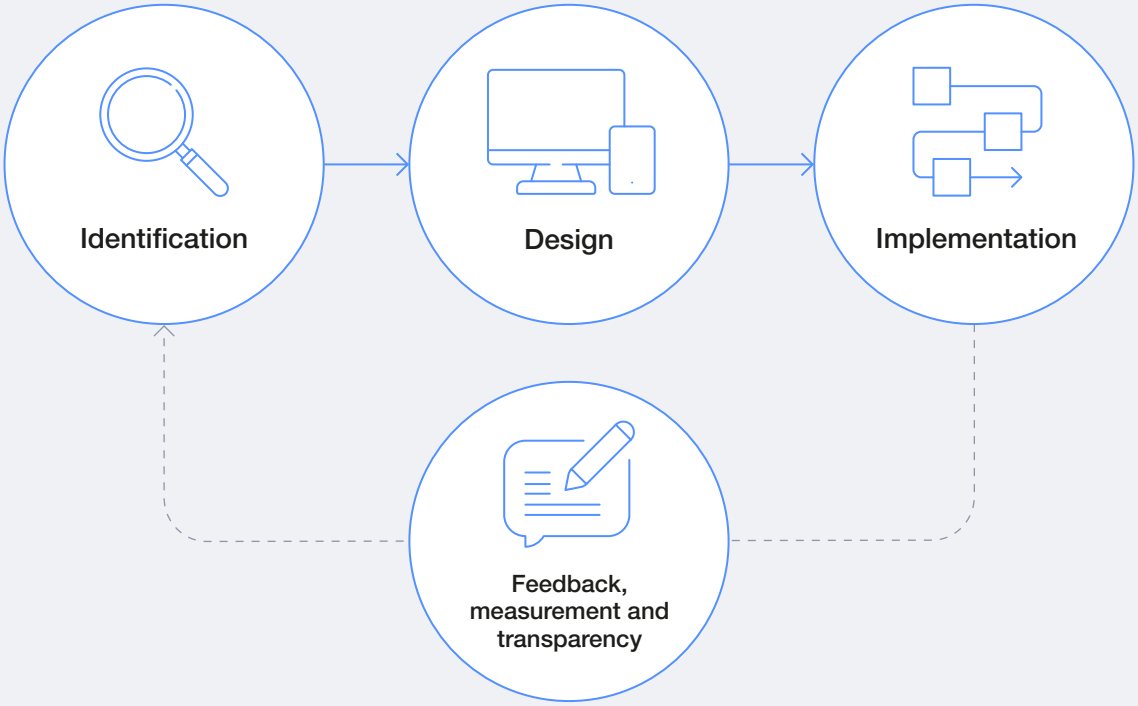
1 Mapping the intervention journey

Embedding Safety by Design principles throughout development creates resilient, adaptable interventions that address evolving digital safety challenges.

The implementation of an intervention is a journey that must be mapped to identify and address the challenges and risks that could arise at each stage of its development and deployment. By mapping the journey, valuable lessons can be learned to help anticipate obstacles such as technical limitations, resource constraints or user adoption issues. This proactive and preventive strategy, actioned at every stage of the journey, follows the principles of Safety by Design. This comprehensive approach not only ensures the intervention's effectiveness but also aids in refining and adapting it to meet evolving demands. Additionally, the mapping ensures that both short- and long-term considerations are woven into the decision-making process.

The intervention journey should be seen not as a linear process but as a series of key considerations that guide the development of interventions. The process is often – and sometimes should be – iterative. These steps in the journey are not exhaustive checklists but rather broad stages that include example considerations to help inform decision-making.

FIGURE 1 **Intervention journey steps**



1.1 Identification

It is necessary to identify risk factors specific to an organization, including characteristics such as service functionalities, user base and business models that can contribute to potential harm. Since there is no one-size-fits-all approach to managing risk, different risk categories may be more relevant to certain types of platforms or services (e.g. social media platforms versus dating apps) or in specific geographic regions. Risk factors cannot be assessed in isolation, as the level of risk may be affected by the cumulative impact of multiple risk factors. Knowing and understanding why certain risk factors exist can help in prioritizing risks.

The identification and mitigation of digital safety risks should be achieved through comprehensive risk assessments. The coalition's report, *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*, introduces a risk assessment framework designed to guide this process, as outlined in Figure 2.

FIGURE 2 | **The identification journey**

Source: World Economic Forum. (2023). *Digital Safety Risk Assessment in Action: A Framework and Bank of Case Studies*, p. 7. https://www.weforum.org/publications/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies/.

Risk assessments must account for the diverse demographics that may be affected, including both users and non-users. This follows the Global Coalition's guidance that digital safety includes preventing harm to non-users: "Online platforms can also be evaluated based on their processes, tools and rules designed to promote the 'safe use' of their services in a manner that mitigates harm to vulnerable non-user groups."[1] As organizations develop and implement these assessments, it is crucial to ensure they are integrated across the entire organizational footprint.

BOX 1 | **Identification**

Key considerations

- What are the geographic, cultural or regulatory factors that could influence risk levels?
- Which groups or segments are most vulnerable to these risks, and have individuals with lived experiences been consulted?
- What functionalities or services in the platform may contribute to potential harms?
- What previous incidents or patterns of harm have been encountered, and what were their causes?
- Are there any external factors, such as ecosystem influences or off-platform risks, that should be considered?
- What resources, partnerships or tools are available to help address these risks?

1.2 Design

When designing digital safety interventions, user protection should be a priority from the very beginning. The principles of Safety by Design ensure that protective measures are woven into the fabric of a product or service. This approach helps prevent potential harms before they arise but also reinforces user trust by embedding safety throughout the entire development process. It is also key to embed safety into the culture and leadership of an organization, cultivating accountability and aiming to create more positive, civil and rewarding online experiences for everyone.[2]

The three principles that Safety by Design is built upon are: service provider responsibility, user empowerment and autonomy, and transparency and accountability.

1. **Service provider responsibility**: Requiring organizations to take proactive steps in anticipating and mitigating potential risks before they impact users
2. **User empowerment and autonomy**: Providing users with clear age-appropriate options and controls over their online experiences, enabling them to make informed decisions about their privacy, security and interactions, and designing features and functionality that preserve fundamental consumer and human rights, including the right to safety
3. **Transparency and accountability**: Involving open communication around the risks, processes and policies associated with a digital product or service, allowing users to make informed choices and accurate assessments of what is working

Accountability, on the other hand, ensures that organizations are responsible for the safety of their platforms and their impact on users and vulnerable non-user groups.[3] Each principle encompasses multiple steps that organizations should take to preserve safety during the development process.

Implementing Safety by Design principles is not without its challenges. One significant obstacle is the potential conflict between safety measures and other business objectives, such as maximizing user engagement or revenue generation. For example, stricter safety controls may limit certain functionalities that are popular among users. Additionally, embedding safety from the outset can be resource-intensive, requiring significant investment in research, development and ongoing management.

Despite these challenges, the importance of Safety by Design cannot be overstated. By embedding these principles into the very core of their operations, organizations can not only protect their users but also build a more resilient and trustworthy digital ecosystem while contributing to brand safety, ad revenue and user retention.

BOX 2 | **Design**

Key considerations

- Is the intervention scalable and adaptable to future risks or growing user bases?
- Is the intervention aimed at prevention, mitigation, detection or response?
- How can the intervention's design balance user privacy with safety measures?
- Are there internal resources (e.g. staff, technology, expertise) to develop and implement the desired intervention?
- How will user trust be built and transparency around how the intervention works ensured?
- How will the effectiveness of the intervention be evaluated over time?

1.3 Implementation

A primary consideration in implementing an intervention is the alignment with the organization's overall digital strategy and goals. This involves ensuring that the intervention is not only technically sound but also compatible with the organization's culture, values and existing processes. Scalability is another crucial factor, as interventions need to be adaptable across various platforms, user bases and geographical regions. Integration with existing systems is also essential, as new safety measures must work seamlessly with current technologies and workflows without causing disruptions or inefficiencies. Overcoming resistance to change within an organization is a key challenge, as employees and stakeholders may be hesitant to adopt new practices, especially if they require significant shifts in how they operate or interact with technology. Addressing this resistance through clear communication and training, and involving key stakeholders early in the process are vital actions for ensuring successful implementation.

Resource planning, the allocation of time and budget, and access to skilled personnel can also significantly impact the success of digital safety interventions. Organizations must allocate sufficient resources to not only develop and deploy these interventions but also to maintain, measure and update them over time. Budget limitations can restrict the scope of what can be achieved, leading to difficult decisions about which safety measures to prioritize. Time constraints can also be a significant challenge, especially when interventions need to be implemented quickly in response to emerging threats. The availability of skilled personnel who understand both the technical and strategic aspects of digital safety is often limited, making it difficult to find and retain the expertise needed to effectively manage these interventions. Proper resource planning, therefore, involves a careful balancing act of ensuring that the organization can meet its safety goals without overextending its capabilities. Digital safety interventions are best seen as essential costs of doing business, with organizations building them into their budgets from the start. Financial, reputational and other business costs of unmitigated online harms may cost a company more in the long term.

BOX 3 **Implementation**

Key considerations

- Has intervention been tested for functionality, security and user impact before a full rollout?
- How will the organization ensure users are informed about the intervention and how it impacts their experience on the platform or service?
- Has the organization identified potential risks or challenges during the implementation phase and does it have contingency plans?
- What mechanisms are in place to monitor and respond to potential unintended consequences?
- Will a potential partner's involvement increase the intervention's scalability, reach or effectiveness?
- Do staff and stakeholders have the necessary training and tools to implement and monitor the intervention effectively?

1.4 Feedback, measurement and transparency

Although listed as the final step in the intervention journey, measurement and review should be integrated into every stage of the process.

To ensure the effectiveness of digital safety interventions, continuous monitoring needs to be established. This ongoing oversight can include automated tools that detect and flag potential safety issues, as well as regular audits to evaluate the intervention's effectiveness and identify areas for improvement. By maintaining a vigilant watch over how interventions operate in practice, organizations can ensure that their safety measures remain effective and relevant in a rapidly evolving digital landscape.

While automated tools can be used alongside other measures to determine the effectiveness and repercussions of an intervention, the user and their feedback are key. An essential component of continuous monitoring is the establishment of clear and accessible channels for user feedback. User feedback enables organizations to identify and address specific issues while offering valuable insights into how safety interventions are perceived and experienced across diverse user communities. It serves as a source of both quantitative and qualitative data. By actively listening to users and incorporating their feedback into the development process, companies can refine their interventions to better meet the needs and expectations of their user base. Organizations should seek feedback from non-governmental organizations (NGOs) or civil society organizations working with those with lived experiences, as they can provide valuable insights that users may generally lack.

Within this step, defining key metrics is fundamental to evaluating the success of digital safety interventions. Metrics could include the number of incidents reported, response times, user satisfaction and awareness of safety features, and overall harm reduction. Measuring digital safety is a complex task, complicated by the constant evolution of technologies, the challenge of developing metrics that are both adaptable and consistent, and the need to balance privacy concerns with the demand for transparency. A previous coalition paper, *Making a Difference: How to Measure Digital Safety Effectively*, categorizes digital safety metrics into three groups:[4]

- **Impact**: metrics that illuminate the impacts on individuals and provide insights into characteristics and patterns of lived experiences
- **Risk**: metrics that enable the detection and mitigation of potential harms
- **Process**: metrics that cover the approach, implementation and outcomes of systems relating to digital safety

The transparency of an organization's digital safety, maintained through regular reporting on safety interventions, is a major pillar of the review and feedback of interventions. Transparency reports allow organizations to share information about the actions they are taking to protect users, the effectiveness of those measures and the challenges they face. By openly communicating about their safety efforts, companies can build trust with their users and stakeholders, demonstrating accountability and a genuine commitment to creating a safer digital environment. These reports also provide an opportunity to showcase progress, celebrate successes and highlight areas where further work is needed, promoting a culture of continuous improvement and collaboration in digital safety efforts.

Although listed as the final step in the intervention journey, measurement and review should be integrated into every stage of the process.

BOX 4 **Feedback, measurement and transparency**

Key considerations

- Are the feedback mechanisms accessible and easy to use for all affected parties?
- How transparent is the organization being with users and stakeholders about the goals, processes and outcomes of the intervention?
- What information can the organization publicly share to demonstrate commitment to digital safety while maintaining user privacy and mitigating bad actors circumventing safety measures?
- How is the organization collecting feedback from users, partners and stakeholders regarding the digital safety intervention?
- Is the organization adequately reporting on both successes and areas needing improvement in the intervention's impact?
- How will the feedback inform future iterations of the digital safety intervention?

1.5 Partnerships

While not a distinct step in the intervention journey, partnerships play a crucial role throughout the entire process. These partnerships can provide valuable insights into the types of interventions most suited to addressing specific challenges, drawing on the diverse experiences and expertise of different organizations and industry experts. Additionally, partners can offer guidance on developing effective metrics for measuring the success of interventions, helping to ensure that the outcomes are both meaningful and actionable.

Harnessing the knowledge and resources of others can allow companies to learn from past experiences and avoid repeating mistakes.

When entering into partnerships, it is necessary for all parties to ensure there is a clear agreement regarding publicity, time requirements, internal and external updates, implementation and other factors. This collaborative approach highlights a broader perspective, enabling organizations to address digital safety challenges more comprehensively and with greater confidence. Furthermore, partnerships can lead to the sharing of best practices, innovative solutions and new technologies, all of which contribute to more robust and effective interventions. In a rapidly evolving digital landscape, the strength of an organization's network of partnerships can be a determining factor in the success of its digital safety initiatives.

BOX 5 **Partnerships**

Key considerations

- Which organizations, platforms or industry experts have successfully addressed similar risks?
- Can the organization effectively address the identified risks without external expertise or resources? If not, what type of expertise, resources or technology does it need from a partner?
- Are there any pre-existing networks, coalitions or multistakeholder groups the organization can join for broader collaboration?
- Can a partner provide critical resources, such as training programmes, to enhance the organization's internal capacity?
- Will a partnership help accelerate the development or deployment of the intervention?
- Could a partnership provide access to new technologies or tools that are currently unavailable?

2 Intervention case studies

Case studies illustrate how structured interventions address risks, refine safety measures and strengthen global digital ecosystems.

The presentation of the following digital safety intervention case studies, organized through the intervention journey steps, is essential for providing organizations with a clear, structured understanding of how to approach and implement effective interventions. By following the intervention journey framework, organizations can navigate the practical nuances and challenges involved at each stage. This step-by-step approach ensures that organizations can customize their strategies to align with their unique needs, resources and risk environments. Moreover, illustrating each step underscores the interconnectedness of the process, highlighting the iterative nature of digital safety mechanisms, which require continuous refinement in response to feedback and evolving threats.

These case studies add significant value by showcasing real-world applications, successes and learnings across diverse contexts and industries. They offer actionable insights and concrete examples that other organizations can model or adapt to strengthen their own digital safety efforts.

Disclaimer: The following examples of case studies in this section are presented within the broader context of digital safety and are intended solely for educational purposes. Their inclusion does not imply endorsement or approval of any specific activities, organizations or strategies discussed. Readers are encouraged to critically evaluate the information and consider it in light of their unique contexts.

2.1 Enhancing gaming safety: a toolkit for parents and guardians

Microsoft developed the *Xbox Gaming Safety Toolkit*, a comprehensive resource for parents and caregivers. The toolkit is designed to be a one-stop shop to help parents and caregivers understand common gaming safety risks and identify the parental controls and user tools available on Xbox. The variety included in the toolkit makes it a technical, behavioural and educational intervention.

The toolkit explores five common risks associated with gaming and provides age-specific guidance on how to respond to these concerns, recognizing that appropriate interventions evolve as the child gets older.

This resource was initially developed with government and civil society partners from Australia and New Zealand and was tailored for these two markets. Microsoft has since expanded the resource to Singapore, Japan and South Korea, and worked with partners in these markets on localization to help ensure the advice and examples are appropriate to each context.

Identification

The development of the toolkit was prompted by recognition of the fact that additional efforts were needed to help raise awareness of existing family safety tools and the role these can play in digital parenting strategies. This featured feedback from a range of external sources and trends in Microsoft's Global Online Safety Survey, including a clear message that parents wanted to ensure their children were able to game safely but felt overwhelmed by negative messages about online risks.

The toolkit focuses on user education, recognizing that parents and caregivers may not have a good understanding of the available tools or how to best deploy those to support their family's specific safety needs. It also recognizes that some families may not be aware of common online safety risks for young people.

Through the convening of a working group to help identify key harm areas to focus on and receive feedback, the toolkit identified five main harms: harmful and age-inappropriate content, bullying and harassment, grooming and unwanted contact, screentime and in-game purchases. These harms correspond to categories outlined in a previous coalition report, such as threats to personal and community safety, harm to health and well-being, and violations of dignity.[5]

Design

The intervention was designed to empower parents and caregivers and help them feel more confident in how and when to deploy tools to support safe gaming for their children.

To apply the toolkit, Microsoft convened the previously mentioned multistakeholder working group with representatives from Australia and New Zealand. These included representatives from the eSafety Commissioner, the Alannah and Madeline Foundation and the Department of Home Affairs in Australia, and Netsafe and the Te Mana Whakaatu Classification Office in New Zealand.

The working group roundtable specifically focused on gathering feedback on the most important harms to profile, and the best way to structure this resource and identify any additional key messages. A key piece of feedback from the group was the need to tailor advice by the age of a child, noting that at different stages of a young person's development,

“ The intervention was designed to empower parents and caregivers and help them feel more confident in how and when to deploy tools to support safe gaming for their children.

90%

of parents use at least one tool to help monitor their child's online activities.

different interventions will be appropriate, and the level of support needed from a parent will change over time. The group also provided important feedback on the need to strike a positive tone of empowerment in the resource to avoid putting caregivers off engaging with the content.

The overall intervention consisted of a combination of both a practical resource and a supporting communications strategy to try and raise awareness with parents. This was important in highlighting key points from the toolkit and encouraging parents and caregivers to engage with the topic of gaming safety more broadly.

For subsequent country-specific iterations, Microsoft has worked with locally relevant policy departments and civil society organizations in Singapore, Japan and South Korea.

Collaboration has been critical in the design of the toolkit – it would not have been successful without input from a diverse range of local experts helping to highlight the different challenges that parents and caregivers may be facing, areas of common misunderstanding and priority online safety risks to discuss in a gaming environment.

Implementation

The launch of the toolkit was accompanied by media campaigns to reach parenting and consumer audiences and raise awareness of the resource (given Microsoft's annual Global Online Safety Survey results show that teens turn to their parents for help after experiencing online safety risks and that 90% of parents use at least one tool to help monitor their child's online activities).[6] Building awareness of the launch was critical in encouraging knowledge-sharing and supporting its uptake.

The Gaming Safety Toolkit itself is structured in three key parts.

1. **Introducing the topic of parental controls and explaining why these should be used as part of a broader family conversation around online safety**: Here, the toolkit encourages parents and caregivers to have open conversations with their children about gaming and involve them in setting rules and boundaries for their own gaming. It also provides a summary of some of the ways that parental controls can be harnessed to set additional guardrails for children online.

2. **Exploring the five key risk areas**: These include harmful and age-inappropriate content, bullying and harassment, grooming and unwanted contact, screentime and in-game purchases. This part provides a summary of how these risks may manifest and ways that caregivers can respond. This section also features a mix of resources relevant to the country where the toolkit has been released. For example, in Australia, the toolkit points to relevant resources from Australia's eSafety Commissioner.

3. **Featuring age-specific advice for carers, broken into four age groups (5-8 years; 9-12 years; 13-15 years; and 16 years and over)**: Within these age groups, the content is broken down into "learn", which outlines common risks for this age group; "explore", covering key tools and conversation starters; and "support", which links to age-appropriate local support resources. These sections also include age-specific case studies illuminating the common risks discussed in the first part of the toolkit.

While some of the tools discussed in the resource are specific to Xbox, the tips are generally applicable to most online services.

Feedback, measurement and transparency

Regarding the measurement of the intervention, Microsoft is still in the process of building its understanding of the toolkit's impact. Measuring the success of educational and behavioural interventions is always a challenging prospect. As outlined in the coalition's paper *Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online*, the difficulty is due to a number of factors including the dynamic nature of technological advances, the diverse array of digital products and services, the concurrent evolution of harmful behaviours, the vast volume of content and the contextual or subjective nature of certain types.[7]

However, as proxy measures, Microsoft has been able to track downloads of the toolkit in different countries, engagement with the communications campaigns and qualitative feedback from stakeholders.

2.2 Safeguarding young gamers: a tailored gaming solution

> k-ID automatically adjusts game settings based on local laws and enables features to be tailored to each child's age or digital maturity level.

k-ID launched a global compliance engine that enables online publishers and platforms to incorporate Safety by Design. Powered by a first-of-its-kind compliance database and on-device facial age estimation technology, k-ID automatically adjusts game settings based on local laws and enables features to be tailored to each child's age or digital maturity level. For families, k-ID developed a cross-platform, unified set of parental management tools to help guide kids and teens anywhere in the world to privacy-preserving, age-appropriate online experiences. This intervention not only focuses on technical safety measures but also incorporates educational and behavioural strategies to empower both parents and children in the digital age.

Identification

k-ID identified several critical online harms that pose significant risks, including threats to personal and community safety, harm to mental health and well-being, exposure to hate and discrimination, violations of privacy, and deception through manipulation.[8] These issues are particularly pressing as children increasingly spend time online, often on platforms that lack age-appropriate safety measures. Recent studies show that 86% of children aged 6-12 have access to digital devices,[9] and 60% of children aged 8-12 use platforms despite age restrictions.[10]

Partnerships with child safety organizations provided insights into the latest trends and risks in online harms, particularly those that are prevalent in social media and games. Many platforms, while claiming to be for users aged 13 and over or 16 and over, are accessed by underage users, often with parental consent to misrepresent their age. Complex and inconsistent definitions of "child" – ranging from 7 to 21 years old across jurisdictions – make uniform compliance challenging. Furthermore, non-compliance is increasingly risky, with fines increasing tenfold in the last three years compared to the previous 20, and the associated costs of investigations and consent decrees being even more burdensome.

The high costs associated with legal, compliance and engineering efforts to create and maintain comprehensive safety systems represent another significant risk. These costs are a major barrier, particularly for smaller platforms that may view these systems as cost centres. Current transaction-based verification processes make large-scale safety checks impractical due to costs, leading to inconsistent verification and significant safety gaps. This underscores the need for scalable, cost-effective solutions that simplify integration across multiple platforms and reduce ongoing engineering costs.

Design

To address these risks, k-ID developed a technical solution that automatically configures game settings based on a child's age and location. This approach also allows parents to tailor their child's online experiences according to their digital maturity, enabling them to gradually introduce more complex or interactive features as their child becomes ready. This is complemented by policy-based interventions, such as compliance with local regulations regarding online child safety and educational strategies aimed at empowering parents and children. k-ID's privacy-preserving, on-device age estimation technology ensures accurate age verification without exporting any biometric data, complying with regulations like the General Data Protection Regulation (GDPR) and the Children's Online Privacy Protection Act and protecting user privacy.

63%

of parents find it challenging to set up and manage parental controls across different devices.

k-ID has continuous engagement with game developers, child safety organizations and leading regulators who provide data and insights. k-ID's technology is certified by the Entertainment Software Rating Board and the Age Check Certification Scheme.

k-ID regularly consults regulators around the world. Regularly consulting with regulatory and industry experts, k-ID aims to harness partnerships to achieve genuine safety improvements rather than superficial compliance.

Co-design collaborations have enabled k-ID to continuously refine its platform based on real-world feedback. For example, game publishers ensure that k-ID remains aware of key compliance issues and offer feedback to ensure seamless integration without affecting gameplay. These partnerships have facilitated access to real-world data on how children interact with online platforms, informing more effective safety interventions.

Implementation

To ensure compliance with global regulations, k-ID monitors and updates the legal landscape across regions. k-ID's comprehensive compliance database, used by industry leaders such as Google, Roblox, Take-Two and Hoyoverse, enables cost-effective maintenance and updates of safety features for gaming companies.

The platform consolidates essential services into a single, cost-effective solution reducing costs associated with monitoring, assurance, verification, regulatory engagement and policy analysis. The revenue-based pricing model ensures even smaller platforms can afford top-tier compliance and safety features, making digital safety accessible and scalable across the industry.

k-ID's architecture is designed for minimal server loads, with most processes occurring at the device or publisher level, enabling easy scalability across gaming environments. Through a partnership with game developer Another Axiom, k-ID launched a pilot integration programme on Gorilla Tag, a virtual reality (VR) game. This test validated the platform's effectiveness and provided a blueprint for scaling.

Feedback, measurement and transparency

Feedback has been gathered through co-design conversations, real-world implementation and continuous user testing with parents and children, safety experts and game developers. While not a trust and safety tool driven by threat-detection algorithms, k-ID ensures transparent feature settings and permissions aligned with validated regulatory frameworks, enabling continuous improvement. During its own research, k-ID found that 63% of parents find it challenging to set up and manage parental controls across different devices, and 81% prefer a simple setup with fewer options. Reflecting these findings, initial feedback from parents indicated a need for clearer communication about how k-ID's safety features work. In response, k-ID redesigned the user interface to include more intuitive explanations and visual aids, leading to a 75% increase in user satisfaction.

At this early stage of k-ID's integration with Gorilla Tag, the focus is on impact metrics to gauge the effectiveness of its platform in enhancing user safety and satisfaction. Impact metrics, one of three categories presented in a Global Coalition for Digital Safety paper, focus on translating subjective user experiences into quantifiable, objective data related to content or conduct.[11] These metrics illuminate user harms or benefits within the digital realm. During the initial beta phase, k-ID served hundreds of thousands of Gorilla Tag players from over 140 countries, delivering age-appropriate experiences based on their age and location. Adjustments ranged from blocking under-10s (due to Meta's requirements) to requiring verified parental consent for ages 10-12 and modifying communication features for teens aged 13-17 in certain jurisdictions. Gorilla Tag is incorporating this feedback in its re-launch, with k-ID fully integrated.

With support from DesignSingapore, k-ID will run co-design programmes with secondary school students, gathering their perspectives and refining onboarding processes and future features. Through these workshops, continuous testing and close collaboration with stakeholders, k-ID will gain valuable insights, highlight areas for development and help shape a future where children can safely play, learn and grow up online.

2.3 Cultivating trust in the digital age: a holistic approach to artificial intelligence and literacy

TikTok's approach to artificial intelligence (AI) and digital literacy is a multi-pronged effort aimed at empowering creators to safely and responsibly express themselves using AI, supporting viewers with important context around AI-generated content (AIGC) and promoting digital literacy. This comprehensive approach entails several components, including AIGC labels, external partnerships, countering harmful misinformation and ensuring that users have access to the rules, tools and skills they need to have a trustworthy experience on TikTok and beyond.

Identification

TikTok's AI and digital literacy interventions stemmed from their goal of creating a welcoming, safe and trustworthy environment. The interventions specifically seek to tackle harms relating to violation of dignity, invasion of privacy and deception and manipulation.[12]

While recognizing that AI offers vast creative potential, it is also important to acknowledge that it can confuse or mislead viewers if they are unaware that content is AI-generated. Without proper labelling, AIGC could more easily spread misleading information or disinformation, making it more pervasive and effective in deceiving users. However, even with the labelling of AIGC or edited media, it still may be harmful.

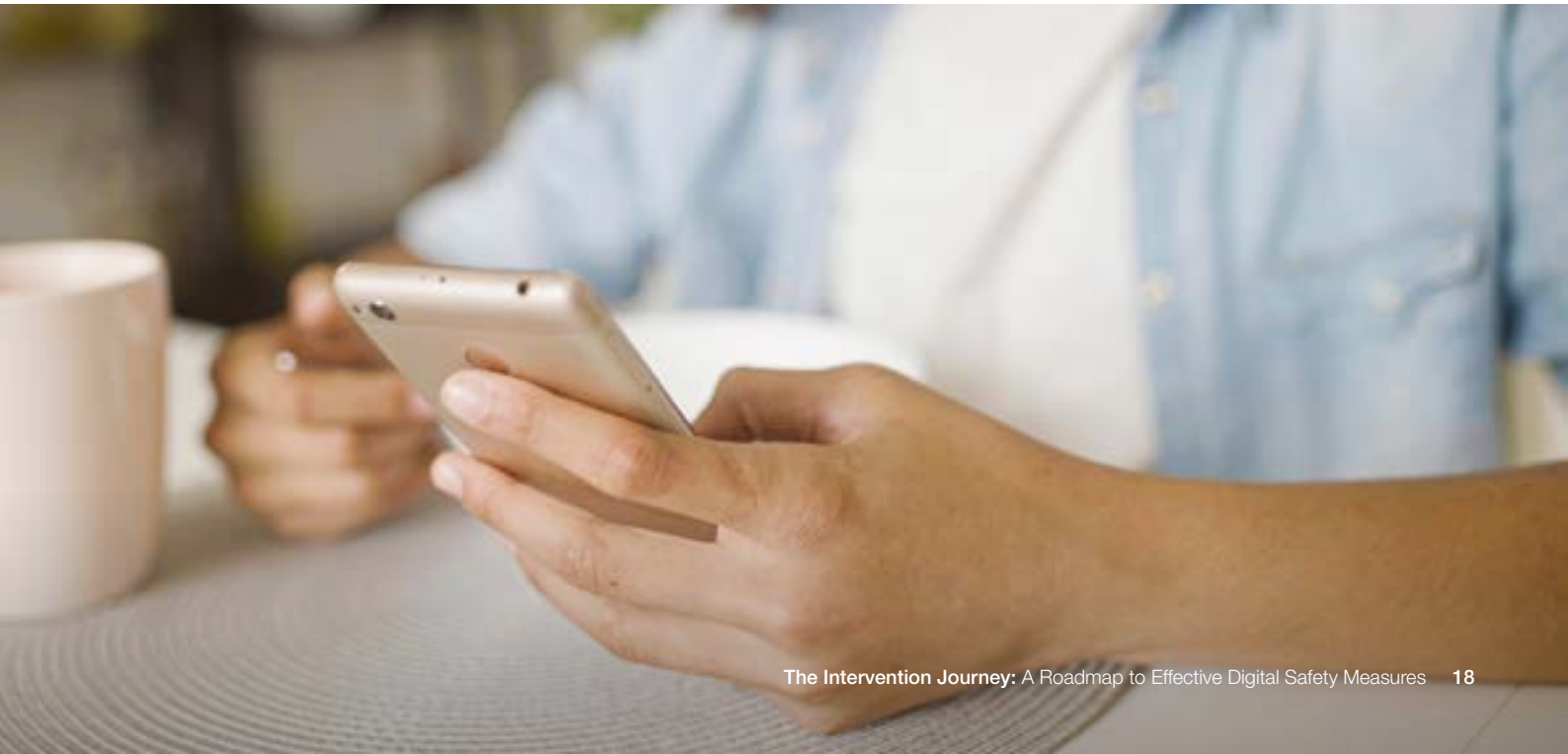
Another driving factor was the threat posed by deceptive actors using AIGC to target online platforms during elections. This is particularly concerning in the context of covert influence operations, where networks of accounts collaborate to mislead users and manipulate platform algorithms. Such actors may attempt to strategically influence public discourse, undermine election results, sway opinions on armed conflicts or shape discussions on social issues. These inauthentic behaviours pose a significant risk to the integrity of online platforms and public debate.

However, even with labelled content and a reduced prevalence of misinformation, it remains essential for users to understand what the labels mean and how to interpret them. General media literacy education is also crucial for cultivating a safer and more trustworthy platform.

Design

TikTok's approach to AI and digital literacy entails a mix of policy-related, technical and educational interventions. Some of these efforts predate the recent rise of generative AI and have evolved accordingly, while others, such as AI content labels, have been directly responsive to this changing digital landscape.

The countering-misinformation strategy involved developing policies that strictly prohibit harmful, misleading AIGC, whether labelled or not. This approach includes creating more proactive AIGC detection models and consulting with experts and industry peers to design effective solutions. TikTok tested methods for automatically labelling AIGC and provided creators with tools to label content that has been fully generated or significantly altered by AI. Specifically, in dealing with elections, TikTok



> Since September 2023, 37 million creators have used the tool that auto-labels AIGC made with TikTok AI effects.

partnered with the AI Elections Accord to set expectations for how to manage the risks arising from deceptive AI content.

In partnership with the Coalition for Content Provenance and Authenticity (C2PA), TikTok enhanced its auto-labelling by incorporating content credentials that attach metadata to content, enabling the instant recognition and labelling of AI-generated material. TikTok also joined the Adobe-led Content Authenticity Initiative to collaborate with a cross-industry ecosystem focused on restoring trust and transparency online. This capability has been implemented for images and videos, with development still under way for audio-only content.

To help TikTok's community understand how to use its tools, the platform launched a video campaign focused on explaining the origins of content on TikTok. Developed in consultation with the international human rights organization WITNESS, the campaign also encourages users to report content that appears AI-generated but lacks proper labelling, raising awareness of the issue. TikTok partnered with MediaWise, a programme of the Poynter Institute, to create videos focused on universal media literacy skills, explaining how tools like AIGC labels can help provide additional context for content. Additionally, TikTok is collaborating with industry peers to support the National Association for Media Literacy Education's new AI Literacy Initiative, aimed at spreading AI literacy to a wider audience.

TikTok also draws essential external advice from its network of Safety Advisory Councils, which include nine regional councils and a US Content Advisory Council. These councils bring together experts from diverse fields, such as youth safety, free expression and hate speech. Their insights play a key role in shaping policies, refining product features and helping TikTok stay proactive in addressing emerging safety challenges.

Implementation

To implement its counter-misinformation strategy, TikTok combines technology and human expertise to combat misinformation at scale. This effort includes specialized misinformation moderators equipped with advanced tools and training, as well as local teams that collaborate with experts to ensure responses consider regional context and nuances. TikTok has also partnered with 19 global fact-checking organizations that assess content accuracy in over 50 languages and help enforce its misinformation policies.

TikTok's endorsement of the International Foundation for Electoral System's Voluntary Guidelines for Election Integrity for Technology Companies is crucial, given the key role platforms play in communicating election information. These guidelines set expectations and practices for companies and election authorities to promote election integrity. By implementing measures to provide trustworthy information to users, TikTok helps maintain public trust in its platform during critical election periods.

TikTok's Edited Media and AIGC policy mandates that creators label AIGC or edited media that features realistic-looking scenes or individuals. This can be accomplished using the AIGC label or by including a clear caption, watermark or sticker. Through the partnership with C2PA, TikTok automatically labels AIGC uploaded from certain other platforms. Additionally, TikTok's policies prohibit content that misrepresents authoritative sources or crisis events, as well as content that falsely portrays public figures in contexts such as being bullied or making endorsements. The platform also disallows the use of likenesses of young people or adult private figures without their consent.

In addition to labelling, implementing a series on media literacy content addressing topics such as misinformation, AIGC and AI transparency helps prevent confusion among viewers who may not understand the meaning behind these labels and also helps them to detect misinformation more easily.

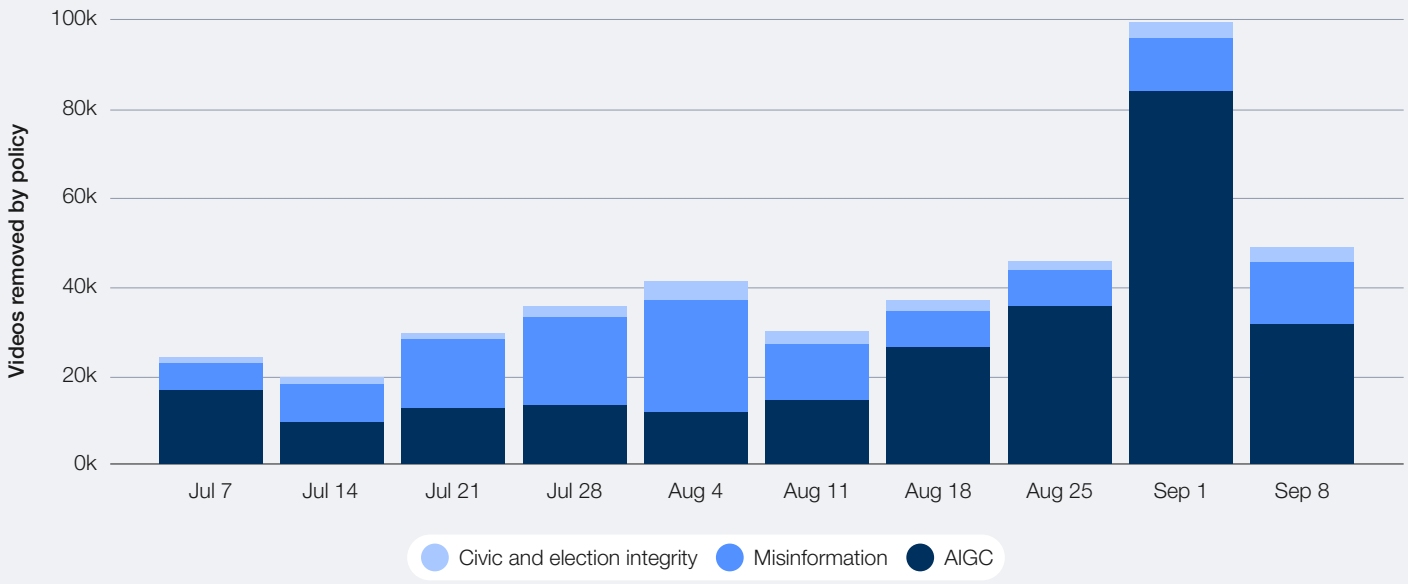
Feedback, measurement and transparency

There are several metrics that offer insight into TikTok's AI and digital literacy efforts. The first is focused on the auto-labelling tool. Between September 2023 and May 2024, 37 million creators have used the tool that auto-labels AIGC made with TikTok AI effects. Relatedly, as the first video-sharing platform to implement content credentials, the increase in auto-labelled AIGC on TikTok may be gradual at first. As other platforms also implement content credentials, TikTok will be able to label more content. TikTok collects and publicly shares data related to community guidelines enforcement. The data is granular, and relevant sub-metrics can be helpful in understanding how TikTok's policies are put into practice. This data includes:

- TikTok's proactive removal rate for civic and election integrity, misinformation and synthetic and manipulated media was nearly 99%, and over 89% of those videos were removed before any views.

- Between the week of 7 July 2024 and mid-September 2024, TikTok has removed over 250,000 edited pieces of media and AIGC that violated their policies.

FIGURE 3 | **TikTok video removals**



Notes: The fluctuation in removal rates of AIGC from 25 August-1 September is related to actions to counter AI cryptocurrency scams, and was not directly related to the US election.

Source: TikTok. (n.d.). *US Elections Integrity Hub.*

2.4 Online safety for teens: protections and supervision features

Meta developed Instagram Teen Accounts, a comprehensive resource for teens and parents, providing built-in protections and new exploration tools. This initiative addresses three key concerns associated with teen online safety: limiting harmful contact, restricting inappropriate content and promoting healthy online habits.

Teen Accounts automatically apply protections to all Instagram teen users (under 18), requiring parental permission for settings changes for users under 16 and offering enhanced supervision features. To ensure accuracy, age verification requirements are implemented, and proactive account identification technology is also being developed.

Identification

Teen Accounts were identified as a method to give parents greater oversight of their teens' experiences. Different ages, situations and scenarios influence parents to set certain rules, and that's why parents need simple, efficient ways to oversee their teens' experiences. Many experts consulted by Meta said that supervision is not an on-off switch, and it is important to meet young people and parents where they are.

Research and consultation informed key concepts in the development of Teen Accounts, starting with research showing how teenagerhood is a transitional time when teens are learning how to express themselves and how building community – whether online or offline – plays an important role in their development. Teens may use online spaces for connection and to discover new things, and the Teen Accounts feature intends to preserve these spaces for them when considering new, protected experiences.

Research with teens also showed that, generally, they want their online experiences to be positive. The teens consulted say they tend to look for continued support from social media apps, including tools that would address situations where they: perceive they are wasting time, come across unappealing content or come into contact with people who make them uncomfortable.

Teens are interested in ways that digital services might provide transparency and controls to help drive age-appropriate experiences, and also see value in their parents having more insights. Teens view supervision tools as useful since their changing patterns of behaviour can be used as evidence to show their parents or caregivers that they are ready to handle greater responsibility.

> With Instagram Teen Accounts, teens are defaulted to the strictest setting of the platform's sensitive content control.

Design

Teen Account protections are designed to address parents' biggest concerns, including who their teens are talking to online, the content they are seeing and whether their time is being well spent. While Teen Accounts put new protections in place automatically, many parents want to be even more involved in their teens' experiences, so the launch of Teen Accounts also comes with new additional elements beyond Instagram's current supervision feature.

Designed with parental support in mind, Teen Accounts are designed to automatically apply baseline protection settings for all teens so parents can feel assured that their teens' experiences on Instagram are protected.

Parents are concerned that their teens might see mature or inappropriate content online, which is why Meta has stricter rules around the kinds of content teens see. Meta's Community Standards and Community Guidelines outline what is allowed and not allowed on Facebook and Instagram respectively and are designed to protect the whole community – including teens – from harmful content.

With Instagram Teen Accounts, teens are defaulted to the strictest setting of the platform's sensitive content control. Teens under 16 cannot change this setting without a parent's permission.

Teens may try to circumvent these new protections, which is why they must verify their age. Since 2022, Meta has required teens to prove their age through a video selfie or ID check if they attempt to change their birthday from under the age of 18 to over 18.

AI technology will be used to predict if someone is over or under the age of 18. Meta trained this technology with signals like profile information, when a person's account was created and interactions with other profiles and content. From those signals, calculations can begin to be made about the likelihood of whether someone is an adult or a teen, even if a teen has listed an adult birthday on their account.

Implementation

Teens who sign up for Instagram are automatically placed into Teen Accounts, and existing teens already using Instagram were notified and shifted into Teen Accounts within 60 days of launch.

Parents decide if teens under 16 can change any of these settings to be less strict:

- **Private accounts**: With default private accounts, teens need to accept new followers and people who do not follow them cannot see their content or interact with them. This applies to all teens under 16 and teens under 18 when they sign up for the app.
- **Messaging restrictions**: Teens will be placed in the strictest messaging settings, so they can only be messaged by people they follow or are already connected to.
- **Sensitive content restrictions**: Teens will automatically be placed into the most restrictive setting of sensitive content control, which limits the type of sensitive content teens see in places like the explore and reels pages.
- **Limited interactions**: Teens can only be tagged or mentioned by people they follow. Hidden words are also activated so that offensive words and phrases will be filtered out of teens' comments and DM requests.
- **Time limit reminders**: Teens will get notifications telling them to leave the app after 60 minutes each day.
- **Sleep mode enabled**: Sleep mode will be turned on between 10pm and 7am, muting notifications overnight and sending auto-replies to direct messages.

If parents want more oversight over their older teens' (16 and over) experiences, they simply have to turn on parental supervision. Then, they can approve any changes to these settings, including:

- Get insights into who their teens are chatting with while not being allowed to read their teens' messages
- Set total daily time limits for teens' Instagram use
- Block teens from using Instagram for specific time periods
- See topics their teen is looking at

The launch of Teen Accounts was accompanied by a launch event with parent creators, as well as media campaigns to reach parenting and consumer audiences. The primary goal was focused on raising awareness of the resource. Ensuring families know about and understand Teen Accounts is key to promoting safer online experiences for teens.

Feedback, measurement and transparency

Meta continuously engages in research and close consultation with academics, parents, teens and other stakeholders to inform the development of safe, age-appropriate experiences. Specifically, the Teen Accounts feature was developed in regular consultation with Meta's Youth Advisors and Safety Advisory Council, which includes third-party experts and professionals in diverse fields such as online safety, privacy, media literacy, wellness and social and emotional health.

In developing Teen Accounts, Meta also consulted with stakeholders to understand their perspectives and inform the ultimate implementation approach. Furthermore, since 2018, Trust, Transparency and Control Labs has consulted with more than 600 stakeholders, 300 teens and 270 parents from more than 350 countries to inform a number of the safety and privacy features of Meta technologies. These consultations have helped develop age-appropriate experiences for teens that preserve their access to online connection and community.

In this process, it has also been key to evaluate external guidance from governmental bodies and children's rights groups.

2.5 Addressing CSEA risks: a chatbot for deterrence and support

A prompt/response chatbot entitled "reThink Chatbot" was developed with the aim of deterring users from searching for CSEA, intervening and directing them to seek support to help change their behaviour. The chatbot is mainly a technical, behavioural, educational and partly policy-related intervention.

Aylo (operator of online adult entertainment platforms) maintains a list of more than 28,000 banned terms in multiple languages, which is constantly being updated. When a search uses a banned term, the chatbot appears as a pop-up, along with a warning message. In short, if users enter a search term associated with child sexual abuse material (CSAM), they 1) receive a warning, and 2) a chatbot operated by Internet Watch Foundation (IWF) appears on their screen. Through the information provided in the warning or by engaging with the chatbot, users are informed about the illegality of CSAM, and they are referred to the Lucy Faithfull Foundation's (LFF) free, anonymous support and advice services, which are provided for people who are concerned about their attraction to CSAM.

> Reducing demand reduces supply and therefore prevents such abhorrent material from being created in the first place.

Identification

The prevalence of CSEA online is an ever-present and growing issue. If interventions can be made early enough, then LFF believes that users can be stopped through intervention. Services that LFF and other NGOs provide around the world include brands such as Stop it Now, Safe to Talk and Talking for Change.

The objective was identified as preventing searches and attempted consumption of CSEA material. Reducing demand reduces supply and therefore prevents such abhorrent material from being created in the first place.

Deterrence messages for those searching for CSEA material and other harmful content were already in place on Aylo's adult entertainment platforms, who worked directly with LFF and similar organizations around the globe to provide the messaging and the local resources to affected users. The IWF and LFF sought to try other ways of signalling to such users to seek help to change their behaviour.

Design

The IWF created a working group to help devise the chatbot and ensure expert input into its development. The group met frequently throughout the months leading up to the release of the intervention to ensure all voices were heard and expertise was considered. Frequent discussions took place within the working group throughout the development of the chatbot. Focus groups were used to evaluate the prompts and responses.

Partners in the design process beyond the IWF also include LFF and The University of Tasmania (UoT). All three organizations provided anonymous use data to UoT, which analysed it in detail and produced a public report on its success. LFF created the prompts and responses for the chatbot and tested them in focus groups.

The combined cross-sector working group proved very successful in ascertaining how the chatbot should function and what metrics were to be analysed. The varied pieces of expertise provided by each party allowed the project to develop at a high standard and achieve its goals.

Implementation

The reThink project aims to minimize the demand for CSEA material and to make users aware of the Stop It Now! support service when they attempt to search for CSAM on a legal pornography website in the UK.

The reThink intervention commenced in February 2021. A warning message was displayed on one of Aylo's online adult entertainment platforms in the UK whenever a user searched for a term that Aylo had designated as potentially relating to an interest in CSAM. This warning message was displayed for 12 months, unchanged.

On 11 March 2022, the warning message was supplemented with a chatbot that had been developed by IWF. The chatbot was a simple conversational agent built on the Google Dialogflow ES platform. The chatbot allowed users to click on buttons to select a path forward to information or to manually enter text. The chatbot functionality was designed to minimize the risk of inappropriate responses, and provided an efficient way for users to quickly receive information about the support services available. Responses are predefined, and not made using generative AI.

The chatbot conversations were monitored throughout the length of the intervention, which enabled it to be continuously improved to ensure it was as effective as possible at connecting users with support services. This included adding additional responses and modifying how the chatbot operated to ensure appropriate responses were provided to users.

The original warning message operated uninterrupted and unmodified for 14 months, after which it was modified twice to trial different wording and to test the effect of removing the chatbot. The warning message was modified on 11 May 2023, and then again on 8 June. The chatbot was disabled from 6 July through to 3 August. Data collection ended on 31 August 2023.

Feedback, measurement and transparency

User feedback was extremely limited due to the type of intervention and the users it was targeting, though responses entered into the chatbot by users are being used for future developmental changes. The entire project was evaluated by a third party – UoT – and their detailed analysis is available online.[13]

In evaluating whether the reThink chatbot was effective, the following metrics were analysed:

- Number of sessions containing CSAM-related searches
- Total number of CSAM-related search queries
- Number of chatbot views per session
- Number of chatbot triggers per session
- Search types after chatbot trigger

- Length of interaction time with chatbot
- Number of LFF URL clicks
- Number of individual visits to helpline
- Number of individual visits to "get help" info pages

As an outcome of the analysis, the warning message and chatbot were displayed approximately 2.8 million times between March 2022 and September 2023. In total, 82% of users only saw a single warning and then desisted from searching for CSAM terms. Interactions with the chatbot by users resulted in 1,656 responses asking for more information and Stop It Now! services, and there were 490 recorded click-throughs from the chatbot to the Stop It Now! website. A total of 68 calls or chats made to the Stop It Now! helpline were identified as likely being prompted by the reThink chatbot and/or warning page on the online adult entertainment platform in the UK. Prior to the chatbot's launch in March 2022, the warning message was displayed an additional 2,208,864 times, totalling 4,400,960 times over the length of the project.

This outcome demonstrates that there has been a clear benefit in the reThink project, as individuals have requested the support of the Stop It Now! service because of the intervention. The evaluation also shows a clear deterrence effect, with a reduction in CSAM search volume on the adult entertainment platform.

However, reThink's impact diminished over the length of the intervention and most markedly after the first three months. The evaluation demonstrated that some users who interacted with the chatbot with typed messages had a negative experience or did not gain the support needed. Future versions of the chatbot should aim to be more complex and less one-size-fits-all, enabling them to be empathetic and adaptive without being therapeutic in unexpected situations.

2.6 Global Signal Exchange: combating online scams and fraud

> Combating scams is challenging because each organization and sector can only be aware of a small portion of the scam life cycle.

The Global Signal Exchange (GSE) is a new collaborative, industry-led project between the Global Anti-Scam Alliance (GASA), the DNS Research Federation (DNSRF) and Google (as first founding member).[14] The GSE aims to become a global clearing house for online scams and fraudulent bad actor signals.

Identification

The core abuse types this project targets are scams and fraud. Scams have been increasing in volume and complexity.[15] They are often carried out by transnational crime organizations; bad actors who operate at scale constantly adapt their methods and combine online and offline activity to lure people into their fraudulent schemes, making it difficult to track these and intervene effectively.[16] The GSE is set up to be expanded to more abuse types, making it extremely fungible and adaptable to future focus areas.

Combating scams is challenging because each organization and sector can only be aware of a small portion of the scam life cycle. In addition, bad actors innovate at a rapid pace, which is actually increasing with the adoption of AI, making traditional abuse fighting ever more challenging. Existing solutions and efforts have been either fragmented, segregated by geography, abuse type or sector, or very narrow in scope and reach, which together make it hard to connect them at a global level and responses less versatile when responding to new types of scams. Google's experience shows that effectively combating scams and the criminal networks behind them requires robust collaboration among industry, businesses, civil society and governments to thwart bad actors and safeguard users. The GSE overcomes these challenges and limitations, allowing the exchange of bad actor signals globally in real time.

Design

The GSE was born out of the partnership between the GASA, the DNSRF and Google. It is designed to be global, cross-sectoral and multistakeholder – a concept that has not existed before. In order to be successful, it relies on broader adoption among key ecosystem players from various sectors including banking, telecommunications and law enforcement.

The collaboration harnesses the strengths of each partner:

- GASA's extensive global network of members and stakeholders
- The DNS Research Federation's robust data platform with over 40 million signals
- Google's experience in combating scams and fraud, as well as it its AI capabilities and funding support for the GSE

Google is providing new funding to support the DNSRF and GASA in launching the GSE. By combining efforts and creating a centralized platform, GSE seeks to enhance the sharing of abuse signals, facilitating quicker detection and mitigation of fraudulent activities across diverse sectors, platforms and services.

The platform's data engine, hosted on the Google Cloud Platform, enables participants to share and access signals contributed by others. Harnessing Google Cloud's AI capabilities, it intelligently identifies patterns and matches signals to enhance detection and response efforts.

Implementation

The intervention is based on three pillars: 1) technical, 2) legal and 3) operational.

The DNSRF, a UK-based non-profit organization, initially developed a data engine that runs on Google Cloud Platform to store and analyse DNS-related data for academic and research purposes. Google learned about this platform and DNSRF via its membership in the GASA in early 2024. Google put it to the test, ran a pilot with bad merchant domains from Google Shopping and, in parallel, developed and signed a first-of-its-kind data-sharing agreement with DNSRF. This has allowed Google to test the platform's capabilities and initiate an internal signal exchange programme to be able to create and manage the exchange of bad actor signals via the GSE.

1) Technical: The intervention is a new global data platform that allows the exchange of bad actor data across sectors and geographies in real time. Participating organizations have the ability to share and receive any type of abuse data to improve their own abuse protections and help detect and prevent scam and fraud campaigns faster.

2) Legal: The intervention relies on a data licence agreement that enables GSE to act as a data processor for various types of data, including personal identifiable information of bad actors and most frequently organized crime groups.

3) Operational: This intervention consists of a signal exchange framework through which Google is able to connect its own relevant teams and efforts to the GSE, allowing the exchange to be performed efficiently at scale and driving down the time-to-live, an industry metric indicating the timeframe between when a scam is detected and mitigated.

Feedback, measurement and transparency

The feedback from users and partners was collected over multiple years, peaking in 2021, indicating that abuse data should be allowed to be shared in order to mitigate the proliferation of scams and fraud. Without a complete picture of the life cycle of a scam, it is hard to pinpoint where it starts, which platforms and services it touches, and which interventions could be effective.

In the initial pilot of the data platform, Google was, for the first time, able to share over 100,000 URLs of bad shopping merchants and, as part of the same test, ingest 1 million scam signals. Over time, Google will expand its footprint of platforms and services connected to the GSE, the number of signals shared and ingested as well as actions taken based on the signals exchanged. The impact is measured through identified bad actors across products that, when detected, would be exposed to actions according to the associated policy violations.

2.7 Cross-platform action: collaboration against online CSEA

> Two of the most pressing dangers online today are the sexual grooming of children and financial "sextortion" of young people.

The Tech Coalition's Lantern is the first cross-platform signal-sharing programme for companies to strengthen how they respond to and combat online child sexual exploitation and abuse (OCSEA). The initiative enables participating companies to securely and responsibly share signals with each other regarding activities identified on their platforms that violate their respective policies against OCSEA.

By participating in Lantern, companies can increase their prevention and detection capabilities, speed up identification of threats, build situational awareness of new predatory tactics and strengthen reporting of illegal activity.

Identification

The Tech Coalition identified critical challenges facing the industry in effectively combating OCSEA. These threats are increasingly complex, with predatory actors continually advancing their tactics to evade detection. In addition, OCSEA threats are often cross-platform in nature, making it difficult for any one company to solve the challenge.

Two of the most pressing dangers online today are the sexual grooming of children and financial "sextortion" of young people. In both of these cases, predators often first connect with young people on public forums, posing as peers or friendly new connections, before directing the young person to other platforms to do things like solicit and share CSAM or coerce payments by threatening to share intimate images with others. This means that predators can harness multiple online platforms in the same offence to target children and avoid detection.

Since these offences span across platforms in a single incident, companies can only see a portion of the harm facing a victim. Recognizing this gap, the Tech Coalition highlights an urgent need for companies to work together in order to uncover the full picture and take proper action.

Design

The Lantern programme was developed over a two-year period, which included a pilot phase to define requirements, identify opportunities and ensure that signal sharing remained proportional to the goal of combating OCSEA. Throughout this process, several Tech Coalition member companies, along with other industry participants, contributed to shaping the programme's design and functionality. To further strengthen the programme, the Tech Coalition partnered with Business for Social Responsibility (BSR) to conduct a Human Rights Impact Assessment (HRIA). This assessment helped evaluate the programme's impact and ensure it upheld fundamental human rights principles throughout the programme design.

The HRIA conducted by BSR acted as a guide for embedding human rights considerations into the Lantern programme's design and implementation. As a result, several key safeguards were established, including the creation of a programme taxonomy for all signals uploaded into Lantern, the implementation of a data retention policy for signals and an application/vetting process for companies interested in joining Lantern. Additionally, the HRIA informed the development of annual training programmes covering human rights, data protection and signal processes, which are updated regularly through continuous collaboration with stakeholders.

During the programme's development, the Tech Coalition engaged 23 external stakeholders, including representatives from child rights groups, victim advocacy organizations, digital rights and privacy advocates, government agencies, academia, survivor groups and institutions addressing other trust and safety challenges. This broad consultation ensured that Lantern was informed by diverse perspectives addressing child-safety-related harms.

Implementation

The results of the pilot were compelling. During the pilot, URLs confirmed to contain CSAM were shared with another company. Using these URLs, the receiving company conducted investigations on its platforms, identifying and addressing behaviours linked to these violations. This led to the removal of more than 10,000 accounts engaged in harmful activities.

Encouraged by these findings, Lantern was formally launched in August 2023 and later announced in November 2023. Since its launch, the Tech Coalition has not only increased the number of companies participating in Lantern but has also advanced the programme through enhanced compliance requirements, operational improvements and ongoing risk mitigation and human rights due diligence.

The Tech Coalition remains responsible for Lantern's risk mitigation, management and oversight. This responsibility includes vetting the eligibility of prospective companies, ensuring compliance with the Lantern agreement and

> Through December 2023, participating companies identified, confirmed and took action on 30,989 accounts for violations of policies prohibiting CSEA.

overseeing the programme's day-to-day operations. To become a Lantern participant, companies must undergo a rigorous application process and compliance review before being legally admitted into the programme.

Signals shared through Lantern are not decisions or conclusions but pieces of information requiring further investigation by receiving companies. Yet, a single signal could be the missing piece that helps safeguard a child.

Importantly, Lantern does not facilitate automated actions on a platform. Participating companies independently assess signals before sharing them in Lantern – likewise, companies must confirm that signals downloaded from Lantern correspond to a policy violation on their platform prior to taking enforcement actions. Companies must also provide additional safeguards, such as a user appeals process.

Signals in Lantern are broadly categorized as content-based or incident-based.

- Content-based signals focus on the shared content related to OCSEA, such as CSAM images or videos, manuals or other illegal content. These signals are typically shared as hashes, URLs or keywords.

- Incident-based signals address violations of OCSEA policies where content may or may not be shared, including minor sextortion, sexual grooming, contact offences or trafficking. These signals are usually shared as account information, critical for identifying cross-platform actors evading detection.

After investigating signals, companies can provide feedback to Lantern on its use and outcomes, in line with their policies and applicable law.

Feedback, measurement and transparency

Several metrics are used by the Tech Coalition in measuring the success of the Lantern programme so far. In terms of participant growth, there has been a steady increase since the launch in August 2023, with 25 companies currently participating.

Through December 2023, participating companies identified, confirmed and took action on 30,989 accounts for violations of policies prohibiting CSEA. In addition, 1,293 individual uploads of CSEA material were removed, and 389 URLs/bulk uploads (meaning a given URL could host numerous pieces of content) of CSEA material were removed. This is in addition to enforcement actions by individual companies for terms of service violations.

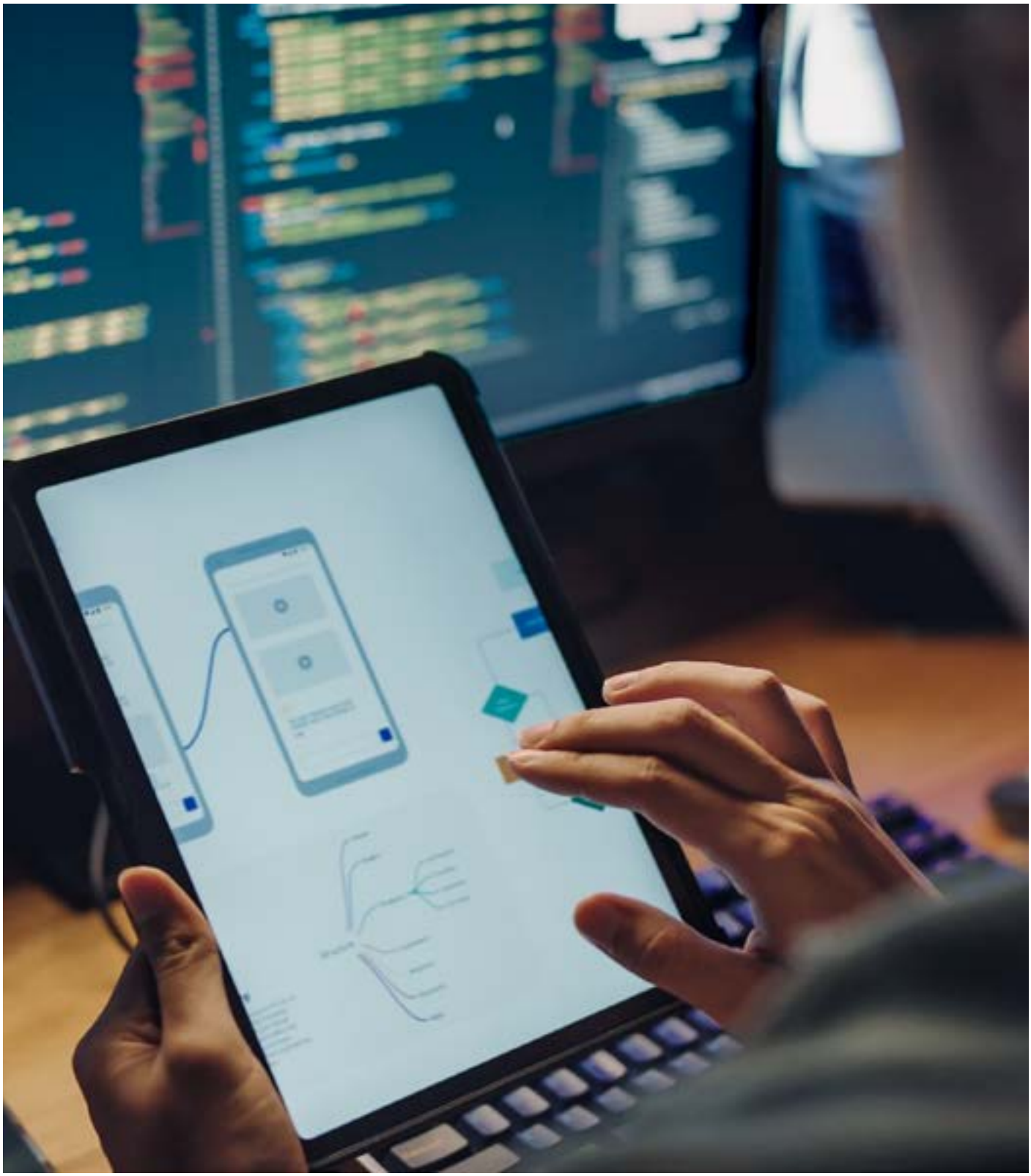
Additionally, 768,044 signals have been uploaded into Lantern.

All information is provided in aggregate, and the outcomes were reported directly by participating companies to the Tech Coalition. A key aspect of this programme is that sharing is voluntary, therefore not all participating companies have shared signals or outcomes. As the programme matures, the Tech Coalition plans to implement ways to increase signal contributions and outcome reporting from participating companies while continuing to consider evolving risk factors.

The Tech Coalition and participating companies will continue to refine the programme and uphold privacy and human rights alongside its mission of protecting young people online.

3 Intervention categories

Digital safety interventions span multiple categories, blending technologies, education, policies and behaviour insights for comprehensive solutions.



> Interventions in the technical category refer to specific technological solutions or tools designed to address and mitigate one of the various forms of online harm.

Identifying the various types of digital safety interventions is essential for creating a holistic response to diverse online harms. Understanding these categories allows stakeholders to tailor solutions that address the specific risks and needs of different user groups, whether through technological tools, awareness campaigns, regulatory measures or behavioural change initiatives. The four categories of interventions are technical, educational, policy-related and behavioural – however, as demonstrated in previous case studies, many interventions do not neatly fit into a single category but instead combine elements from multiple types. This overlap highlights the interconnected and multifaceted nature of interventions needed to tackle varied and dynamic harms.

This section of the report provides more details on each intervention category, along with examples that illustrate interventions within those categories. Many examples encompass aspects from multiple intervention types but have been classified based on their primary feature. It is important to note that these examples are not intended to be an exhaustive list.

BOX 6 **Interventions categories**

Technical: Uses technologies, engineering solutions or technical approaches for measurable and tangible changes

Educational: Disseminates knowledge and enhances skills through structured learning programmes or other informational resources

Behavioural: Centres on understanding, modifying and/or rewarding behaviour through diverse strategies

Policy-related: Focuses on developing, modifying or enforcing policies and regulations

3.1 Technical interventions

Interventions in the technical category refer to specific technological solutions or tools designed to address and mitigate one of the various forms of online harm. This process includes detecting and assessing online harms. Technical interventions provide broad applicability across different types of harms, making them versatile tools. However, a significant challenge lies in effectively grouping harm types and determining when and how a single intervention can address multiple threats.

TABLE 1 **Example technical interventions**

Intervention	Description	Example organization(s)
AI-powered content labelling systems	AI-powered systems that classify content into different categories (e.g. misinformation, hate speech, explicit material), allowing platforms to flag, rate or limit the distribution of harmful content.	**YouTube** (Google) Uses AI to label content as safe or harmful, influencing recommendations and the visibility of videos.
Real-time AI and human-in-the-loop moderation systems	Hybrid content moderation systems where AI flags content and human moderators make the final decision, improving moderation accuracy for nuanced or context-sensitive content.	**Twitch** Identifies potential violations of community guidelines using a combination of automated detection and user reporting.
Nudity detection systems	AI-driven technology harnessing on-device machine learning to detect and blur (when turned on) nudity images. Functions in secure, encrypted environments and may default to active protection for younger users, with prompts encouraging broader adoption among adults.	**Instagram** (Meta platforms) Leverages AI/ML built into the app on user devices. When turned on, it automatically blurs images containing nudity that are being sent and/or received in Instagram direct messages. Teens have it turned on by default, adults need to opt into the feature. Directing people to safety tips when sending or receiving these images, developed with guidance from experts about the potential risks involved.
User identity verification	Technologies to verify users' identities using methods such as facial recognition, ID verification or behavioural analytics.	**Tinder and Hinge** Uses identity verification tools to ensure users are real, reducing catfishing and online abuse.
Blacklisted URL systems for illegal content	Systems that block access to known URLs containing illegal content such as CSAM, preventing sharing or viewing of such content.	**IWF** Works with internet service providers and tech companies to block URLs hosting CSAM.
Hashing technology for illegal content detection	Technologies that use hashing algorithms to detect and prevent the spread of illegal content, such as CSAM or terrorism-related media.	**PhotoDNA** (Microsoft) A hash-matching technology that can be used to detect previously identified CSEA material.

Source: YouTube Official Blog. (2023). *Our approach to responsible AI innovation*; Twitch. (n.d.). *Safety at Twitch*; Instagram. (2024). *New Tools to Help Protect Against Sextortion and Intimate Image Abuse*; Match Group. (n.d.). *Safety*; Internet Watch Foundation. (n.d.). *URL List*; Microsoft. (n.d.). *PhotoDNA*.

3.2 Educational interventions

Educational interventions in digital safety consist of structured programmes, activities or other informational resources designed to inform and educate individuals and groups about digital risks, safe online practices and cybersecurity measures. The goal of these interventions is to enhance users' awareness, knowledge and skills for navigating the digital world safely.

Educational interventions are crucial in raising awareness and providing targeted knowledge. These interventions must build on existing knowledge, progressing through different levels – similar to a school curriculum – rather than starting from scratch. Capacity building for educators is equally important, requiring investments in training and resources to equip them to handle sensitive topics. Furthermore, harnessing technology such as large language models (LLMs) can enhance the learning experience by offering safe, simulated environments where individuals can explore scenarios without real-world risk, facilitating practical and secure learning opportunities.

TABLE 2 **Example educational interventions**

Intervention	Description	Example organization(s)
Media information literacy programmes	Educational programmes designed to teach users (especially young people) how to critically assess media, understand online risks and recognize misinformation.	Meta Collaborated with experts to launch Get Digital, which blends ready-to-use lessons, tips and resources.
Community workshops	Interactive sessions, often hosted in person or virtually, to educate communities about online risks and safety practices, offering hands-on guidance.	Google "Be Internet Awesome" workshops focused on teaching children and parents how to navigate the internet safely, avoid scams and use digital tools responsibly.
Resource libraries – blogs and forums	Online platforms hosting a wide variety of content such as articles, FAQs, guides and expert advice on how to stay safe online and manage digital risks.	Google Safety Center Blog Provides regular updates on online safety practices and privacy tips.
School-based digital safety curricula	Educational curricula integrated into schools to teach students about the importance of digital safety, privacy and responsible online behaviour.	Common Sense Media's Digital Citizenship Curriculum Used in schools across the US to teach students about online privacy, safe social media use and responsible digital behaviour.
Digital well-being resources	Resources aimed at promoting healthy digital habits, reducing screen time and managing online activities to prevent burnout or negative mental health impacts.	Google's Digital Wellbeing App Offers tools to track screen time, set limits on app use, and promote better digital balance.
Parental guides	Resources aimed at improving parents' understanding of how to keep their child safe online and/or use parental controls.	Niantic Parent Guide and Parent Portal to learn about safety tools and features and manage children's experiences online.

Source: Meta. (n.d.). *Digital literacy*; Google. (n.d.). *Empowering kids to be safe, confident explorers of the world*; Google Keyword. (n.d.). *Safety & Security*; Common Sense Education. (n.d.). *Digital Citizenship Curriculum*; Google. (n.d.). *About*; Niantic. (n.d.). *Niantic Parent Guide*.

3.3 Policy-related interventions

Policy-related interventions encompass the development, implementation and enforcement of rules, guidelines and regulations set by companies and governments to address and mitigate online harms.

At the company level, policy-related interventions involve crafting both internal and external policies to address online safety and conduct. Internal policies, for example, include procedures for managing and escalating issues, such as protocols for reporting and handling incidents of harassment or security breaches. External policies, on the other hand, involve policies such as the development of community guidelines that govern user behaviour on the platform.

At the governmental level, policy-related interventions involve enacting and enforcing regulations that set legal requirements for online safety, such as data privacy laws or regulations targeting harmful online content. These policies often require collaboration between different stakeholders.

Developing industry-relevant policies that address the needs of the entire sector, rather than being specific to individual platforms, promotes consistency and broader applicability, such as with policies on storage duration and sharing practices.

Clear processes and policies for handling appeals, empowering content moderation and determining when to remove content are essential for ensuring transparency and fairness in enforcement.

TABLE 3 | **Example policy-related interventions**

Intervention	Description	Example organization(s)
	External	
User appeal process	A formal process enabling users to challenge and request a review of content moderation decisions, such as removals or account suspensions.	YouTube Allows users to appeal video takedowns and account suspensions, reviewed by human moderators.
Policy enforcement mechanisms	A framework that imposes penalties for violating platform policies, such as issuing warnings, strikes, temporary suspensions or permanent bans.	Discord Informs users when they have broken rules, what actions have been taken and how it may impact their overall account standing
Equity and inclusion strategies	Policies aimed at ensuring platform policies and moderation practices protect vulnerable and marginalized groups from disproportionate harm while promoting inclusivity.	TikTok Implements policies to protect minority groups from hate speech and discriminatory content, promoting an inclusive online community.
External engagement with experts and civil society	Collaborating with external organizations, such as researchers, NGOs and civil society groups, to improve policies and content moderation practices and understand risks related to online harms.	Meta Oversight Board An independent panel of experts that reviews content moderation decisions and advises on policy improvements.
	Internal	
Risk intelligence capabilities	Internal systems and capabilities designed to detect, assess and mitigate emerging risks, such as cyber threats, misinformation and other harmful content, guiding platform adjustments.	Google Uses sophisticated risk detection systems to monitor threats like misinformation, harmful content and child safety risks, allowing for proactive policy changes.
Assurance and audit processes	Reviews and audits of content moderation practices to ensure compliance with platform policies, legal standards and ethical guidelines, maintaining consistency in enforcement.	Meta The independent Data Transparency Advisory Group provides an assessment of Facebook's content moderation.
Risk triage and escalation processes	Processes used to assess, categorize and escalate cases of online harms (e.g. terrorism, CSAM) based on severity, ensuring the most serious issues are handled by appropriate teams, including legal and trust and safety teams.	YouTube Uses risk escalation for sensitive content like terrorism or CSAM, where specialized teams make final decisions.
Employee support systems, including for trust and safety teams and others	Platforms that provide mental health support, counselling and stress management resources for employees, especially those exposed to harmful content, like trust and safety and content moderation teams.	Concentrix Monitor psychological safety of moderators and provide ongoing training and practical resources.

Source: Google Transparency Report. (n.d.). *Appeals*; Discord. (2024). *Discord Warning System*; TikTok Safety Center. (2025). *Inclusion and Belonging Guide*; Oversight Board. (n.d.). *Improving how Meta treats people and communities around the world*; Google Safety Centre. (n.d.). *Content safety*; Meta. (2019). *An Independent Report on How We Measure Content Moderation*; Google Transparency Report. (n.d.). *Featured policies*; Concentrix. (n.d.). *Safe Spaces: How to Create Psychological Safety in Content Moderation*.

3.4 Behavioural interventions

Behavioural interventions focus on changing individuals' and groups' actions and habits to reduce digital risks. These interventions apply strategies from psychology and behavioural science to promote safe online behaviours and discourage risky or harmful activities.

Tailoring interventions to fit the specific situation of the individual ensures that the support provided is appropriate and effective. Organizations must develop the capability to implement behavioural interventions by applying approaches that support the effective design and deployment of such strategies.

Additionally, balancing privacy concerns with the need to report potential radicalization to law enforcement poses a challenge. It is important to navigate these concerns carefully, ensuring that interventions are both effective in addressing risks and respectful of individual privacy rights.

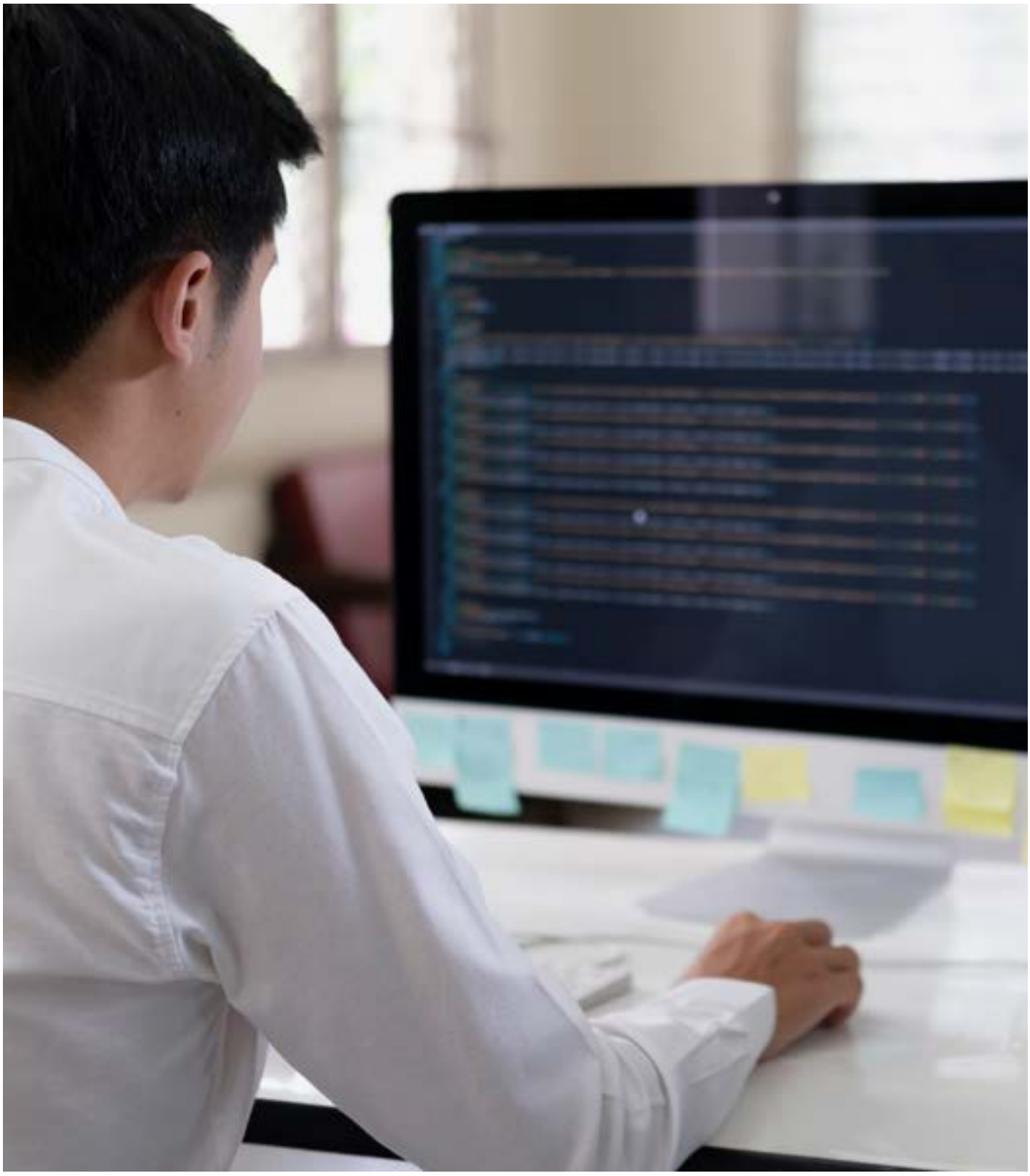
TABLE 4 **Example behavioural interventions**

Intervention	Description	Example organization(s)
Threat modelling mentality	An internal company approach that focuses on understanding how systems and policies can be exploited or "gamed" by malicious users, and designing defences against such exploitation.	Niantic Trust and safety red team exercises to identify gaps or potential problem areas within the product before launch.
Service design	Designing platform features that allow users to control their experience by enabling options such as disabling comments, blocking users or setting privacy filters – this can lead to bad actors ceasing their offensive behaviour by denying them the desired impact.	Instagram Allows users to turn off comments on posts, restrict unwanted interactions and block specific accounts to minimize exposure to harmful content or harassment.
User identification (e.g. detect repeat offenders)	Systems designed to identify repeat offenders or malicious actors using behavioural patterns, unique identifiers or account verification, helping detect banned users trying to rejoin.	Facebook Uses behavioural pattern recognition to detect suspicious activity resulting is disabled accounts and the creation of new accounts under different identities.
Detection of behavioural signals (e.g. keywords/ phrases)	Algorithms that monitor for specific behavioural signals or keywords associated with harmful content, such as harassment, hate speech or threats, to prevent and mitigate harm.	Tinder "Are you sure" feature is a warning to users to think twice about their opening line, using AI to detect harmful language and alerting the sender their message may be offensive.
Positive reinforcement for safe behaviour	Platforms that reward or encourage positive online behaviours, such as respectful interactions, and highlight users who contribute to the platform's safety and positive culture.	Reddit Karma system rewards users with "karma" points for positive contributions to discussions, promoting respectful behaviour and discouraging trolling or abusive behaviour.
Behaviour-focused warnings	Displaying warnings based on user actions, such as sending alerts if a user is engaging in harmful behaviour, or escalating enforcement (e.g. temporary suspension) after repeated offences.	YouTube Provides warnings when users violate content policies, offering them the chance to change their behaviour before harsher penalties like account suspension are imposed.

Source: Niantic. (2023). *Our Approach To Safety*; Instagram Help Centre. (n.d.). *Managing Your Privacy Settings*; Meta. (2024). *How enforcement technology works*; *Tinder Newsroom*. (2021). *Tinder Introduces Are You Sure?, an Industry-First Feature That is Stopping Harassment Before It Starts*; Reddit. (2024). *What is karma?*; YouTube Help. (n.d.). *Community Guidelines strike basics on YouTube.*

4 The SME challenge

Limited resources and expertise often hinder SMEs from implementing digital safety interventions, highlighting the need for scalable solutions.



The digital landscape is not solely full of large multinational corporations – it includes a vast number of SMEs that are equally in need of robust digital safety interventions. However, unlike their larger counterparts, SMEs face additional and often more pronounced challenges. These challenges range from financial limitations, resource constraints, knowledge gaps, organizational barriers and regulatory requirements.

Large corporations face similar challenges, such as allocating proper funding for digital safety and overcoming organizational barriers. However, these challenges are significantly more pronounced for SMEs. SMEs operate in a more competitive environment – competing not only with other SMEs for market share and investment but also with larger, well-established companies. In such a competitive landscape, digital safety investments can be seen as counterproductive to growth, leaving both the companies and their users more vulnerable to digital harms.

4.1 The challenges

Financial

Financial challenges for SMEs in implementing digital safety interventions are significant and multifaceted. One of the primary issues is the high cost of online safety tools and infrastructure. Advanced solutions, such as encryption software, firewalls and threat detection systems, often come with steep price tags that exceed the budgets of many smaller companies. In addition to purchasing these tools, ongoing maintenance, software updates and licensing fees can strain limited financial resources. Employing dedicated staff for digital safety is also a financial burden, as it requires personnel to manage and develop technical systems, as well as handle threat detection, investigation and response.

Resource

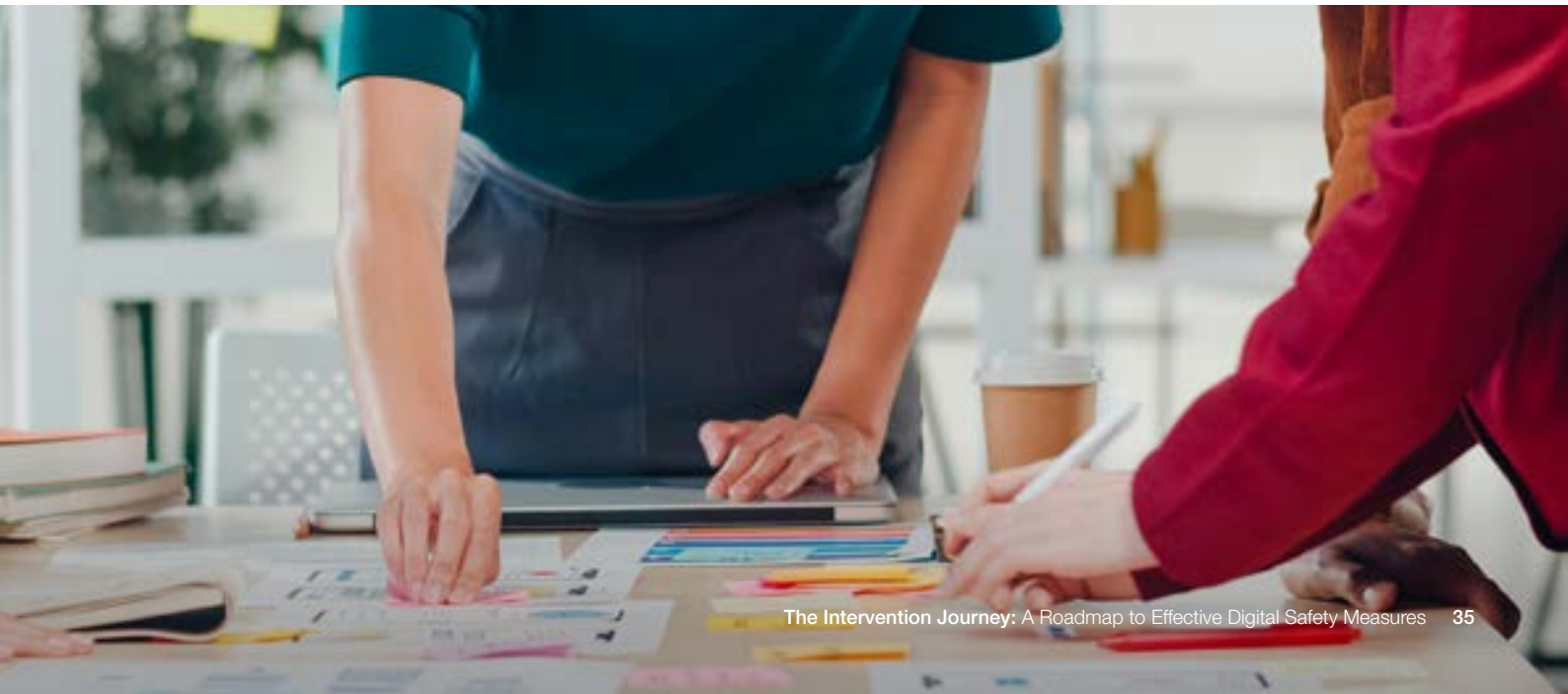
SMEs face significant resource constraints when implementing digital safety interventions, primarily due to their limited access to financial, human and technological resources. Many SMEs operate with small teams, often without a dedicated trust and safety team. This means that employees, already stretched thin with day-to-day operations, are tasked with managing digital safety measures, often without the specialized skills or knowledge required. As a result, critical tasks like monitoring for threats, updating software or conducting security audits may be overlooked or inadequately performed.

Time is another critical resource that SMEs often lack. With limited personnel, there is little bandwidth to devote to researching emerging threats, understanding new regulations or keeping up with best practices in digital safety.

Knowledge

Hiring or developing in-house technical expertise in digital safety is another major hurdle for SMEs. The specialized knowledge required to implement robust digital safety systems is scarce and expensive, making it difficult for smaller businesses to compete with larger organizations that can offer more competitive salaries and development opportunities. Furthermore, the dynamic nature of digital safety – with new threats and vulnerabilities constantly emerging – means SMEs often lack the capacity to provide continuous training for their staff.

“ The digital landscape is not solely full of large multinational corporations – it includes a vast number of SMEs that are equally in need of robust digital safety interventions.



Organizational

Cultural and organizational challenges also play a critical role in hindering the prioritization of digital safety among SMEs. In many cases, these organizations are focused on immediate operational concerns, such as revenue generation and market competition, and may view digital safety as a non-essential overhead rather than a critical business priority. The lack of a safety-first mindset is often exacerbated by leadership's limited understanding of the scope and scale of potential digital threats. Without a clear sense of the risks involved, decision-makers may postpone investments in digital safety or delegate the responsibility without providing the necessary support or resources.

Regulatory

SMEs also face regulatory challenges when implementing digital safety interventions. The increasing complexity of data protection laws and other internet regulations can be overwhelming for small businesses with limited legal expertise or resources. Compliance with regulations such as the GDPR, the Digital Services Act in Europe or similar laws elsewhere requires both financial investment and a thorough understanding of legal obligations. Additionally, regulatory requirements often change, further complicating SMEs' ability to stay compliant without significant external support or legal consultation.

4.2 Solutions and considerations

> No organization should introduce a feature unless they are prepared to effectively manage its associated risks.

Free open-source resources

Providing free resources that help SMEs extend digital safety measures beyond major organizations is critical for tackling digital harms. Instead of starting from scratch, SMEs can harness existing reports and resources as a foundation for their intervention strategies. This approach provides them with a starting point and guidance, easing their journey towards effective digital safety measures.

For instance, the Tech Coalition offers free resources through its Pathways programme to help SMEs understand how companies can combat online child safety exploitation and abuse on their platforms. This includes guidance on external policies, content standards and reporting CSAM to local authorities, such as the National Center for Missing & Exploited Children.

Partnerships

Partnerships offer effective solutions to many of the challenges SMEs face in implementing digital safety interventions. Collaborating with external organizations, industry groups or government bodies allows SMEs to benefit from established interventions and resources without having to develop everything in-house. By joining forces with partners, SMEs can access best practices, technical guidance and tools that are otherwise out of reach.

These partnerships can be tailored to address specific harms or industry requirements, spanning the development and implementation of an intervention, ensuring they are both scalable and effective. Additionally, partnerships can help SMEs navigate regulatory challenges by providing them with expert advice and resources to stay compliant with evolving legal frameworks.

Organizational and technical decisions

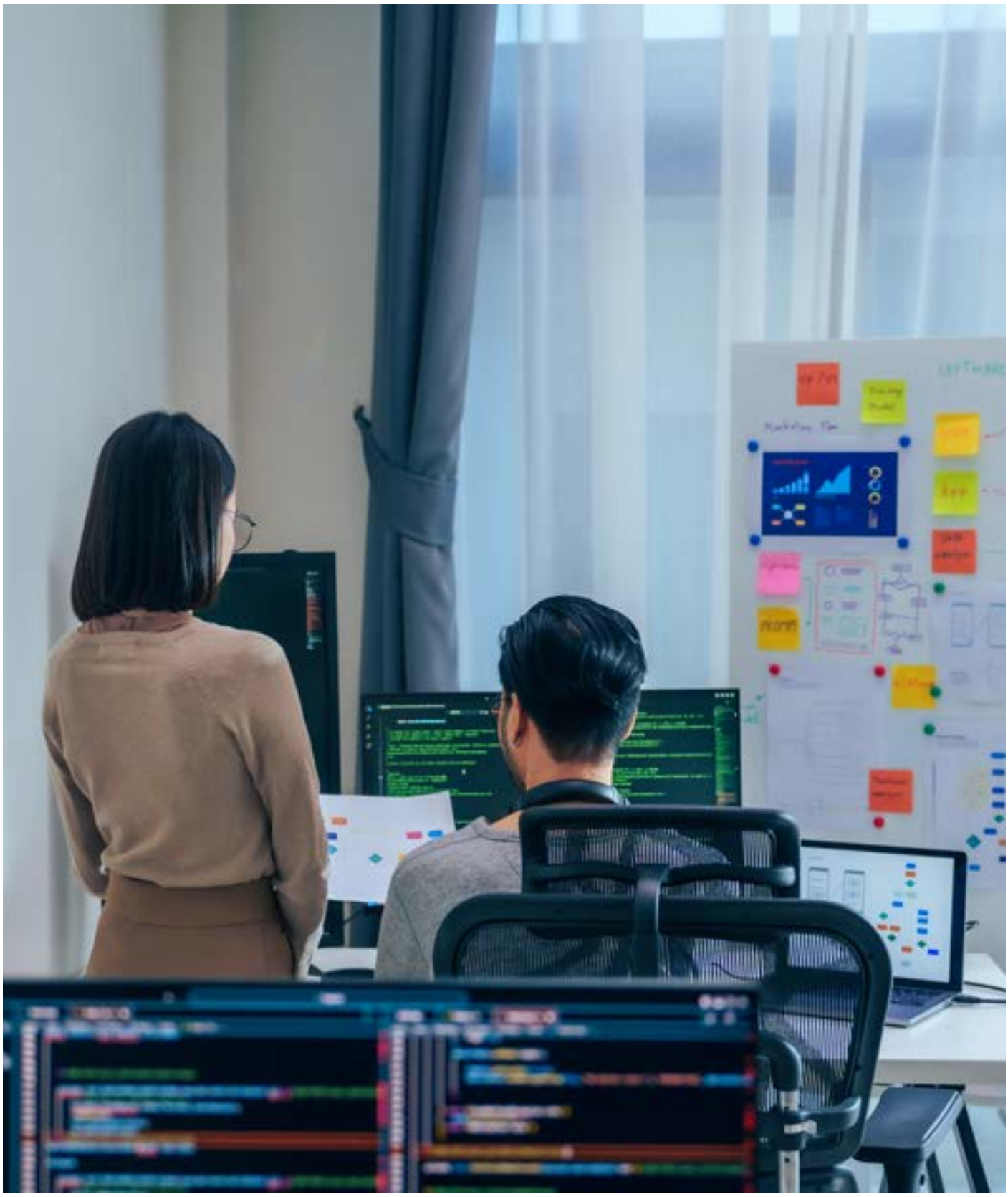
SMEs may avoid adding features like video due to the additional time and resources required to manage the potential harms associated with such features. It is important for these enterprises to carefully consider the risks involved before implementing new features. No organization should introduce a feature unless they are prepared to effectively manage its associated risks.

An SME needs to consider its organizational maturity and development stage. The type and level of intervention that SMEs can realistically implement will differ from those available to larger organizations. Not all SMEs have the capacity to execute high-resource or highly technical interventions.

Interventions should be tailored to their specific needs and capacities. This means creating scalable solutions that can be implemented with limited resources and technical expertise. Starting with basic, effective measures and gradually building their capacity will help improve digital safety practices over time.

5 Recommendations

Addressing complex digital risks requires tailored interventions, responsive designs and collaboration with industry peers and NGOs.



Building on the insights and case studies presented, this section offers targeted recommendations and key considerations for organizations striving to implement effective digital safety interventions. As highlighted, the journey can be challenging, with organizations facing financial, organizational, regulatory and cultural barriers. General strategies to help navigate these complexities are provided here, informed by successful interventions and best practices.

5.1 Adopt a proactive and transparent approach

Organizations should cultivate a culture of accountability, where the responsibility for user protection is primarily borne by the platform rather than solely by the user.

It is crucial to embed safety features proactively, promoting user protections within digital products from the outset to ensure a safer online experience.

As previously mentioned in the report, embedding user protection directly into the digital product and services from inception can preemptively address potential harms. Such features include proactive risk assessments, user-friendly privacy settings and robust content-filtering mechanisms. This proactive design ethos encourages technology companies to shift from a focus on rapid innovation or profit maximization to one that emphasizes intentional, safe and thoughtful growth.

Organizations should cultivate a culture of accountability, where the responsibility for user protection is primarily borne by the platform rather than solely by the user. This entails not only addressing risks during design and implementation but also building a mindset where user safety is a foundational value across leadership and organizational culture. The approach also facilitates inclusive design and consultation of diverse user groups to ensure that safety mechanisms are accessible and effective for all, especially those at higher risk of harm.

Transparency and accountability enable users to make informed decisions. By openly sharing how they enforce safety policies and assessing the effectiveness of their interventions, companies build user trust and set industry standards. Reporting on safety outcomes and sharing innovations that successfully reduce harm encourage widespread adoption of best practices and educate users on proactive steps they can take. This openness not only reassures users but also promotes an ecosystem where digital safety measures are continuously refined and strengthened across platforms.

5.2 One size does not fit all

Interventions should be customized to meet the unique needs of different platforms and user demographics for more effective and inclusive digital safety.

A one-size-fits-all approach to digital safety is often ineffective, as users interact with platforms and face online risks in unique ways. Tailored safety interventions address the specific needs of diverse user groups, platform types and digital ecosystems. For instance, younger users may benefit from enhanced parental controls and age-appropriate content restrictions, while professional networking platforms might focus on protecting user data and preventing impersonation or fraud.

Effective digital safety extends beyond technical measures and requires a holistic approach that incorporates policy improvements, user education and behavioural reinforcement. Non-technical initiatives, such as awareness campaigns, training programmes and digital literacy resources, equip users with the knowledge to navigate online environments responsibly and recognize potential threats. Strengthening internal and external policies can also reinforce these efforts by establishing clear guidelines and protocols for safety.

5.3 Ensure inclusivity in digital safety interventions

It is necessary to adapt protections for diverse vulnerabilities and contexts, offering tailored resources and tools for equitable user protection across regions.

Designing protections that account for users' diverse and intersecting vulnerabilities is essential, incorporating factors such as gender, age, economic status and cultural context. Inclusive digital safety recognizes that certain groups may be disproportionately targeted or impacted by digital harms and adapts interventions accordingly to provide equitable protection. This requires a localized approach, tailoring safety features and resources to fit specific regional contexts, especially in areas with lower digital literacy. Inclusive interventions comprise consideration of the capacity and needs of children and the impact that interventions may have on the full range of their human rights. By implementing inclusive and context-sensitive interventions, platforms can better serve the full spectrum of their user base.

5.4 Implement effective reporting and response mechanisms

Providing accessible reporting tools with strong response protocols ensures timely action and dedicated support for vulnerable users.

In any intervention, accessible, user-friendly reporting tools allow individuals to quickly flag harmful content or behaviours, from cyberbullying to explicit content, enabling platforms to respond swiftly to potential threats. To support this, it is vital that platforms offer clear instructions on how to report issues and what users can expect once they have submitted a report.

Behind these tools, strong response protocols are critical for ensuring that flagged issues are addressed promptly, fairly and transparently. By establishing rapid-response teams that specialize in handling various types of digital harm, platforms can act quickly to investigate and resolve reports in a manner that respects the affected individuals. Additionally, special response protocols should be in place for vulnerable groups, such as children, marginalized communities or victims of gender-based violence, who may experience heightened risks and unique challenges. Escalation pathways for these cases ensure that appropriate and sensitive interventions are available, supporting the protection and well-being of those most at risk.

At a fundamental level, organizations need to conduct reporting and act against reports of illegal content and activities.

5.5 User education and digital literacy

Equipping users with the knowledge of how to use safety tools through tutorials, resources and partnerships cultivates a culture of proactive security.

Educating users about interventions and additional mechanisms is essential for enhancing user engagement and digital literacy. Equipping individuals with the knowledge and skills to navigate these tools effectively can help users to better protect themselves and make informed decisions. The use and combination of interactive tutorials, in-app messaging, accessible resource hubs and educational campaigns can help with this education. For instance, platforms can provide guided onboarding for new features, use timely notifications to prompt safe practices and offer dedicated sections with frequently asked questions and tutorials in multiple languages. Partnerships with digital literacy organizations further broaden outreach, delivering workshops, webinars and public campaigns to raise awareness. Feedback mechanisms and gamification, such as quizzes and rewards, can incentivize users to actively learn and adopt safer practices, building a culture of digital literacy and resilience.

5.6 Prepare for emerging threats

Monitoring evolving digital trends is crucial for anticipating and addressing new risks, ensuring interventions stay aligned with technological advancements.

As the digital environment advances, preparing for emerging threats requires a proactive, future-looking and adaptable approach to safety. Future-proofing interventions involve staying vigilant to new technologies, trends and behaviours that could introduce fresh challenges. This involves closely monitoring shifts in user behaviour, technological advancements and potential vulnerabilities associated with new features, particularly as platforms increasingly incorporate AI, virtual reality (VR) and other technologies. By identifying these potential threats early, digital platforms can prevent risks from escalating and be better equipped to integrate new protective measures before issues arise.

5.7 Strengthen partnerships and multistakeholder collaboration

Collaborating with NGOs, academics and industry peers strengthens the exchange of insights, the development of best practices and the creation of a unified approach to digital safety.

As can be seen through this report, establishing partnerships and collaborating with diverse organizations is essential to creating effective, sustainable digital safety solutions. Collaborating with NGOs, academic institutions and specialized organizations brings a wealth of expertise from fields such as digital safety, child protection, human rights and social sciences.

Industry-wide collaboration among digital platforms is equally crucial for addressing systemic issues that span across the internet. No single organization can tackle all digital challenges alone, nor are organizations competing on digital safety, as harmful behaviours often cross platform boundaries and manifest in various forms. Sharing knowledge, best practices and tools allows platforms to create a unified approach to online safety, standardizing interventions that can be more effective when implemented consistently and widely among organizations.

Multistakeholder partnerships encourage transparency and accountability, with platforms not only benefiting from shared insights but also committing to collective goals that prioritize user safety and digital integrity. Through such partnerships, the digital ecosystem becomes more resilient, cooperative and responsive to the ongoing challenges.

“Educating users about interventions and additional mechanisms is essential for enhancing user engagement and digital literacy.

Conclusion

The intervention journey is both complex and increasingly essential in a rapidly evolving digital landscape. This report encourages a holistic understanding of digital safety, promoting the design of interventions that are user-centric and adaptable to various operational contexts, from large-scale corporations to SMEs. For all organizations, particularly those with limited resources, understanding these key elements of the intervention journey is crucial for tailoring practical and effective solutions.

Achieving comprehensive digital safety is rarely straightforward. It requires a nuanced blend of technical and non-technical solutions to address diverse and ever-evolving digital harms. Interventions need to be adaptable, scalable and context-sensitive to meet the unique needs of each organization. The inclusion of media literacy empowers users to navigate an intervention (and digital spaces) more safely. The importance goes beyond addressing immediate digital harms – it equips users with critical thinking skills that drive resilience and informed decision-making in the face of disinformation and other digital harms.

While digital safety interventions tend to focus on the biggest organizations, SMEs are an essential element of a comprehensive approach to digital safety, as these organizations often face distinct challenges due to limited resources and expertise. SMEs often serve as access points to various services. Ensuring that SMEs have access to effective, affordable digital safety resources is vital to protecting their user bases and maintaining the broader safety of the digital ecosystem. By equipping SMEs with accessible tools and guidance, they are not only better prepared to safeguard their own environments but also to contribute to a network of digitally secure communities. This requires collaborative effort across sectors to provide SMEs with the necessary support to implement robust safety measures that are scalable and tailored to their specific needs.

The use of partnerships is pivotal in the journey for all organizations. Multistakeholder cooperation allows organizations to share best practices and tackle complex issues that surpass the capabilities of any single entity. As digital threats become more sophisticated, the need for coordinated efforts across sectors grows, with coalitions and networks playing a critical role in advancing digital safety.

The challenge of AI in digital safety and the risks it introduces to content moderation and information dissemination are major focus areas. AI-driven tools are increasingly relied upon to identify and mitigate harmful content, yet the ethical and practical complexities of AI-based interventions require careful consideration. Looking to the future, it is crucial for organizations to consider the potential harms that AI could introduce, particularly as these tools become more autonomous and integrated into user interfaces. Future intervention strategies should prioritize the development of AI-specific safety measures that address current challenges while anticipating and mitigating future risks, setting a proactive precedent for ensuring safety by design.

The digital safety intervention journey is one of ongoing adaptation, anticipation, collaboration and innovation. At the same time, the competitive environment of business can result in underinvestment in digital safety from both large and small companies. This report's objective in mapping the intervention journey and using case studies was to make the steps clearer as well as provide general guidance on how to ensure the necessary efforts are taken in digital safety. The Global Coalition for Digital Safety's effort to cultivate multistakeholder engagement on issues of digital safety, such as the intervention journey, is one of the many areas of focus. The coalition will continue to bring together experts and stakeholders to address new challenges as they emerge, amplifying the impact of safety interventions and cultivating a culture of shared responsibility.

Contributors

Lead authors

Daniel Child
Manager, Industry Engagement and Enablement, eSafety

Daegan Kingery
Specialist, Digital Safety and Trustworthy Technology, World Economic Forum

World Economic Forum

Agustina Callegari
Project Lead, Global Coalition for Digital Safety, World Economic Forum

Daniel Dobrygowski
Head, Governance and Trust

Cathy Li
Head, AI, Data and Metaverse; Member of the Executive Committee

Acknowledgements

This paper is a combined effort based on numerous interviews, discussions, workshops and research. The opinions expressed herein do not necessarily reflect the views of the individuals or organizations involved in the project listed below.

Sincere appreciation is extended to the following working group members, who spent numerous hours providing critical input and feedback on the drafts. Their diverse insights are fundamental to the success of this work.

Henry Ajder
Founder, Latent Space Advisory

Maria Cristina Capelo
Head, Safety Policy, Meta

Kay Chau
Vice-President, Programs and Member Success, Tech Coalition

Jeffrey Collins
Director, Trust and Safety, Amazon Web Services

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

Kieran Donovan
Co-Founder and Chief Executive Officer, Kidentify Corp

Sasha Havlicek
Chief Executive Officer, Institute for Strategic Dialogue

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

Adam Hildreth
Founder, Crisp, a Kroll business

Julie Inman Grant
eSafety Commissioner, Office of the eSafety Commissioner, Australia

Heidi Kempster
Deputy Chief Executive Officer, Internet Watch Foundation

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

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

Hayley van Loon
Deputy Chief Executive Officer, Crime Stoppers International

Victoria Nash
Director, Associate Professor; Senior Policy Fellow, Oxford Internet Institute, University of Oxford

Susan Ness
Non-resident Senior Fellow, The Atlantic Council's Europe Center, The Atlantic Council

Bo Viktor Nylund
Director of Innocenti Global Office of Research and Foresight, UNICEF

Jenna Omassi
International Policy Manager, Online Safety, Office of Communications (Ofcom)

Kavya Pearlman
Chief Executive Officer and Founder,
X Reality Safety Intelligence

Katherine Sandell
Platform Risk Program Lead, Trust & Safety, Google

Noam Schwartz
Chief Executive Officer and Co-Founder,
Activefence

Natalie Shoup
Safe Online Specialist, Industry & Data Lead,
Safe Online

Ian Stevenson
Chief Executive Officer, Cyacomb

David Sullivan
Executive Director, Digital Trust
and Safety Partnership

John Tanagho
Executive Director, IJM's Center to End Online
Sexual Exploitation of Children

Liz Thomas
Director Public Policy, Digital Safety, Microsoft

Deborah Welsh
Executive Manager, Strategy, Engagement
and Research, eSafety

David Wright
Chief Executive Officer, SWGfL

John Zoltner
Global Lead, Protection from Digital Harm,
Save the Children International

Production

Louis Chaplin
Editor, Studio Miko

Rose Chilvers
Designer, Studio Miko

Laurence Denmark
Creative Director, Studio Miko

Endnotes

1. World Economic Forum. (2024). *Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online*, p. 20. https://www.weforum.org/publications/making-a-difference-how-to-measure-digital-safety-effectively-to-reduce-risks-online/.
2. eSafety Commissioner. (2024). *Safety by Design*. https://www.esafety.gov.au/industry/safety-by-design.
3. eSafety Commissioner. (2024). *Safety by Design*. https://www.esafety.gov.au/industry/safety-by-design.
4. World Economic Forum. (2024). *Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online*, p. 4. https://www.weforum.org/publications/making-a-difference-how-to-measure-digital-safety-effectively-to-reduce-risks-online/.
5. World Economic Forum. (2023). *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*, p. 7-11. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.
6. Microsoft. (2024). Global Online Safety Survey 2024, p. 38. https://news.microsoft.com/wp-content/uploads/prod/sites/40/2024/02/Microsoft-Global-Online-Safety-Survey-2024.pdf.
7. World Economic Forum. (2024). *Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online*, p. 5. https://www.weforum.org/publications/making-a-difference-how-to-measure-digital-safety-effectively-to-reduce-risks-online/
8. World Economic Forum. (2023). *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*, p. 7-11. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.
9. Dezuanni, M. et al. (2023). *Manifesto for a Better Children's Internet*. Australian Research Council. https://digitalchild.org.au/research/publications/working-paper/manifesto-for-a-better-childrens-internet/.
10. Ofcom. (2022). *Children's Online User Ages Quantitative Research Study*. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/keeping-children-safe-online/childrens-online-user-ages/children-user-ages-chart-pack.pdf?v=328540.
11. World Economic Forum. (2024). *Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online*, p. 4. https://www.weforum.org/publications/making-a-difference-how-to-measure-digital-safety-effectively-to-reduce-risks-online/.
12. World Economic Forum. (2023). *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*, p. 8-11. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.
13. Scanlan, J. et al. (2024). *reThink Chatbot Evaluation*. The Internet Watch Foundation. https://www.iwf.org.uk/media/pvvnnjvf/rethink-chatbot-evaluation-report.pdf.
14. Storey, A., & Zebarjadi, N. (2024). *The new Global Signal Exchange will help fight scams and fraud*. The Keyword. https://blog.google/technology/safety-security/the-new-global-signal-exchange-will-help-fight-scams-and-fraud.
15. Richardson, L., & Adkins, H. (2024). *How to spot scams, and what to do if you encounter one*. The Keyword. https://blog.google/technology/safety-security/how-to-spot-scams-and-what-to-do-if-you-encounter-one.
16. Richardson, L. (2024). *A new way we're helping others track frauds and scams online.* The Keyword. https://blog.google/technology/safety-security/a-new-way-were-helping-others-track-frauds-and-scams-online.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

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

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

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

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