WORLD ECONOMIC FORUM

Making a Difference: How to Measure Digital Safety Effectively to Reduce Risks Online

WHITE PAPER

JUNE 2024

Images: Getty Images, Midjourney

Contents

Foreword

David Sullivan
Executive Director,
Digital Trust & Safety
Partnership

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

Daniel Dobrygowski
Head, Governance and Trust,
World Economic Forum

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

Ensuring a safe online environment is crucial in an age when digital platforms are the cornerstone of communication, commerce and community. But achieving this goal, and measuring progress towards it, is challenging. Dynamic technologies, diverse digital services, rapidly changing harms and evolving regulations must be navigated. Yet despite these challenges, this is a journey worth undertaking.

Sound metrics and measurements lay the foundation for a safer digital ecosystem. They promote accountability, aid evidence-based decision-making, guide resource allocation, facilitate benchmarking and progress monitoring, promote transparency and engagement, and enable the effectiveness of interventions to be evaluated.

This white paper, a publication from the Global Coalition for Digital Safety, reflects the current state of the most salient approaches to metrics and measurements in the online environment. It represents an extensive collaboration by a diverse range of stakeholders, including platforms, regulators, safety providers and members of civil society, academia and international organizations. Over two years, we have convened this expert multistakeholder group to work together on how to measure digital safety.[1]

As an output of this work, the paper proposes grouping the metrics we have identified into three categories: impact, risk and process. This categorization is designed to clarify their application and facilitate stakeholders' tracking and reporting of these crucial aspects. From assessing the effectiveness of interventions to monitoring outcomes, the insights within these pages are invaluable for anyone committed to improving digital safety on their services.

We also hope this contribution will inform and inspire further collaboration across stakeholder groups, promoting ongoing efforts to enhance digital safety outcomes for all.

Executive summary

Effective measurement of online safety informs decision-making, policy development and awareness-raising.

In an increasingly interconnected world, it is essential to measure digital safety in order to understand risks, allocate resources and demonstrate compliance with regulations. However, the task of measurement is fraught with challenges, including ever-evolving technologies, the need for flexible yet consistent metrics and the balancing of privacy concerns with transparency. These dynamics, combined with the fact that the context and nature of harms varies significantly across platforms and types of services, have resulted in an absence of comparable and agreed metrics.

This Global Coalition for Digital Safety paper on metrics and measurements outlines the coalition's current view of the most salient approaches to metrics and measurements in the field of digital safety, drawing on a multistakeholder group comprising platforms, regulators, safety providers, non-governmental organizations (NGOs), academics and international bodies. It provides a structured approach to understanding and evaluating digital safety by promoting a shared understanding of metrics among stakeholders.

Effective management is essential to ensure the risks are properly identified, mitigated and monitored over time. Building on the iterative process for assessing and addressing digital safety risks set out in the World Economic Forum Global Coalition's digital safety risk assessment framework (the risk framework),[2] this new paper categorizes digital safety metrics into three groups:

- **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

Practical application of these metrics is crucial for assessing current safety measures, guiding future improvements and enabling accountability for digital services. Metrics and measurements for digital safety must align with the goals and challenges of the digital landscape.

Recognizing the significance of diverse datasets in digital safety, stakeholders must collaborate on data access while addressing privacy and security concerns as emphasized by the risk framework. Streamlining access and promoting partnerships between researchers and data custodians can enhance data availability.

Continuous improvements in safety measures and increased accountability for digital safety are vital for encouraging a safer online environment. The practical application of digital safety metrics is essential to evaluate interventions and their real-time effectiveness. Digital safety metrics reinforce accountability, empowering NGOs and regulators to oversee service providers effectively. They also serve as benchmarks for compliance monitoring, enhancing user trust in platforms, provided they are balanced with privacy considerations and take into account differentiation among services. Digital service providers should focus on which metrics can be most impactful rather than reading the following recommendations as an exhaustive list of options to adopt.

Measuring online safety enables informed decision-making, facilitates policy development and enhances stakeholders' awareness of online safety issues. In this scenario, regulators can focus on harmonizing benchmarks in order to avoid multiple onerous requirements that would hinder progress towards collecting and evaluating relevant metrics across industry. In addition, this paper recognizes that metrics are just one tool to measure online safety, and should not be considered the only important tracker that can be used to understand progress towards protecting users against online harms.

Introduction

This paper proposes metrics to measure digital safety and provide empirical evidence to guide decision-making.

“ Establishing baseline metrics and tracking key performance indicators over time enables organizations to identify trends and areas for enhancement.

Measuring digital safety outcomes is crucial in an interconnected world for several reasons.

First, it helps organizations grasp the diverse and evolving risks faced by individuals, communities and societies in digital environments. This in turn enables organizations to understand their exposure to threats and vulnerabilities.

Second, measuring digital safety supports effective risk mitigation by allowing for more efficient resource allocation. By identifying and quantifying impacts and risks, platforms and stakeholders alike can develop targeted interventions to achieve improved digital safety outcomes.

Moreover, with new regulatory requirements emerging, measuring digital safety helps organizations demonstrate compliance and assess the effectiveness of interventions for improving safety outcomes by, for example, reducing risks or increasing user empowerment. Additionally, showcasing strong safety measures promotes trust among users, customers and partners, demonstrating a commitment to protecting online users while also minimizing the risk of harm to non-users or the public caused by misuse of platforms.

Finally, establishing baseline metrics and tracking key performance indicators over time enables organizations to identify trends and areas for enhancement. This approach ensures that safety measures evolve alongside technological developments and emerging threats.

Despite its importance, measuring the effectiveness of digital safety interventions comes with challenges. These include the dynamic nature of technology and online harm, the need for consistent yet flexible metrics and the balancing of privacy considerations with transparency. All of these difficulties can be summarized as the challenge of “measuring the immeasurable”.

This Global Coalition for Digital Safety paper on metrics and measurements aims to tackle these challenges head on. It takes a structured approach to thinking about metrics and measurement as they pertain to digital safety. Emphasizing the crucial role of measurement in ensuring good governance within the digital ecosystem, the paper aims to establish a shared understanding of digital safety metrics among diverse stakeholders and promote a common language and mindset in relation to digital safety, especially for smaller or less scrutinized services.

The metrics in question serve as specific indicators or parameters that can be used to assess various dimensions of digital safety, encompassing aspects such as the prevalence of harmful content or behaviours, platform compliance with existing regulations and the effectiveness of safety measures. These metrics offer tangible data points that facilitate comparisons over time or across different entities, enabling a deeper comprehension of trends, patterns and areas for improvement in digital safety initiatives.

> The approach in this paper is intended to promote consistency and alignment across jurisdictions, companies and stakeholders, ultimately enhancing collective efforts to safeguard online spaces.

Measurement involves systematically applying these metrics to collect, analyse and interpret data. This process uses methodologies, tools and techniques to gather pertinent information, evaluate the performance of digital safety initiatives and make well-informed decisions grounded in empirical evidence.

This paper, developed through a collaborative, multistakeholder process, represents a concerted effort to construct a roadmap for understanding those metrics and measurements relating to digital safety. Its structure reflects this collaborative approach. Section 1 addresses the current landscape and challenges of digital safety measurement; Section 2 provides a framework for conceptualizing metrics and measurements; and Section 3 delves into practical applications of online metrics.

By categorizing metrics into "impact", "risk" and "process", the paper facilitates effective risk assessment and intervention design in various scenarios. It takes a holistic approach, recognizing the need to go beyond assessing individual risks to consider societal risks and understand differences in the nature of risks and how to address them. As the Global Principles on Digital Safety highlight, digital safety is about preventing and reducing harm, including through moderating illegal or harmful content or conduct, driving responsible platform design and governance, or designing tools to empower users to tailor their online experiences.[3] This approach is intended to promote consistency and alignment across jurisdictions, companies and stakeholders, ultimately enhancing collective efforts to safeguard online spaces.

Importantly, this paper does not aim to list all possible metrics and measurements in online safety exhaustively. Instead, it seeks to provide a common approach for thinking about metrics relevant to the field.

Finally, this paper is intended to enhance and complement parallel outputs from the Global Coalition for Digital Safety. These outputs include the *Digital Safety Risks Assessment Framework in Action* report,[4] which provides a comprehensive methodology for stakeholders to map out online safety risks, and the *Typology of Online Harms* report,[5] which offers foundational taxonomies for discussing digital safety. Additionally, it aligns with the coalition's Global Principles on Digital Safety,[6] emphasizing the importance of a rights-based perspective on digital safety and advocating for evidence-driven approaches.

1 Context and challenge of measuring digital safety

Frequent changes in technology, regulation and online behaviour make measuring digital safety difficult, but stakeholders are engaging with the challenge.

1.1 The state of play

Public authorities worldwide increasingly recognize the importance of holding online platforms accountable for maintaining safety standards and protecting users' rights.

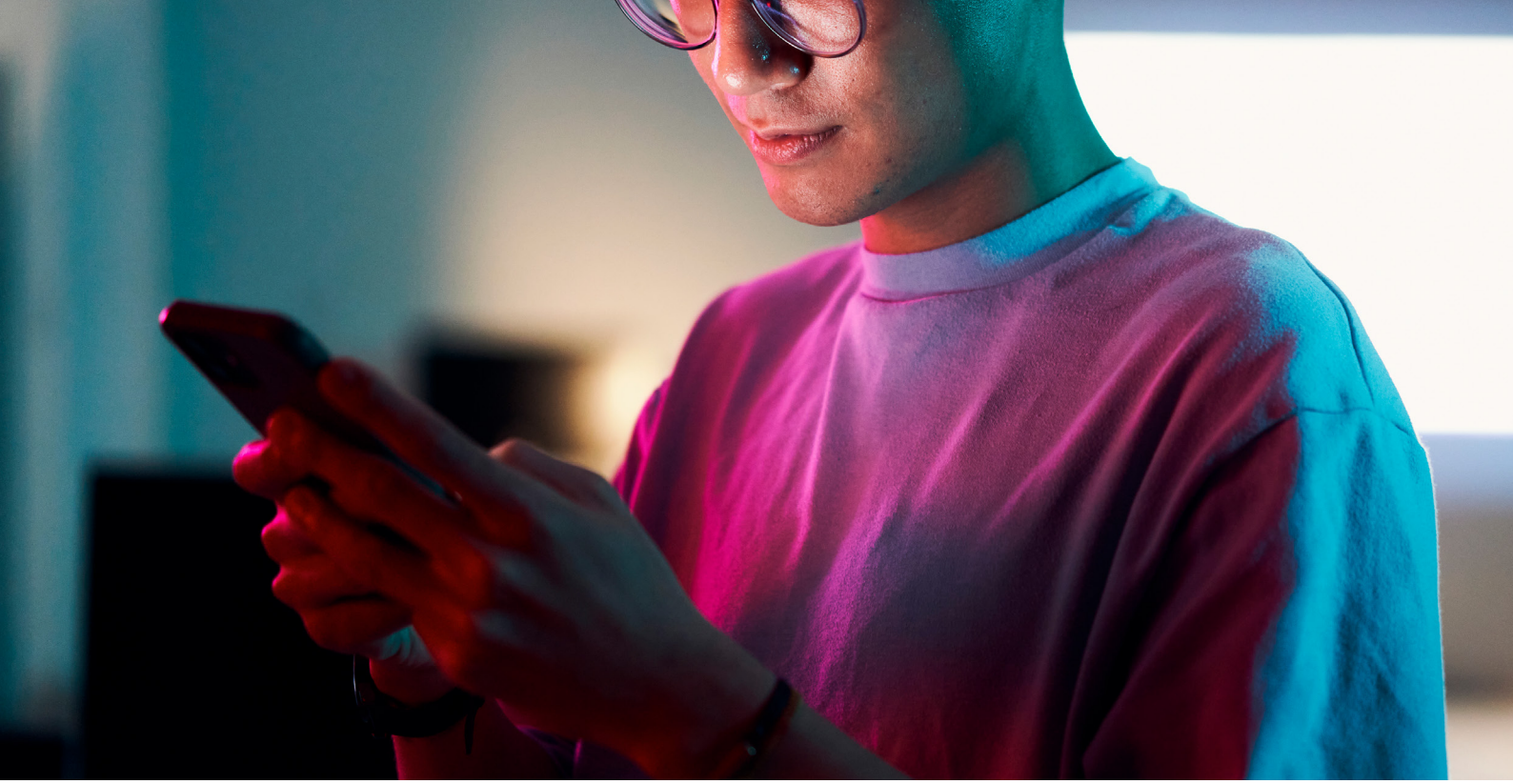
The measurement of digital safety is constantly evolving, shaped by technological advances and regulatory interventions. Technological progress has enabled the development of sophisticated algorithms capable of detecting and mitigating various forms of online harm, such as child sexual abuse material (CSAM), hate speech and disinformation. However, this progress also presents challenges: malicious actors continuously adapt their tactics to evade detection and weaponize new technologies to perpetuate harm such as with generative AI, emphasizing the need for ongoing innovation in online safety measures.

Stakeholders from various backgrounds have for decades championed the call for increased transparency and the disclosure of metrics, exemplified by the widespread growth of initiatives such as platform transparency reporting. As it is known today, the practice of transparency reporting started in 2010 when Google published its first report focusing on government requests for user data and content removal.[7] Some years later, a shift occurred towards reporting on the enforcement of moderation as transparency reporting became more widespread.

These transparency efforts play a pivotal role in promoting trust and accountability within the digital environment. Many services, particularly the most prominent platforms, have proactively collected a spectrum of metrics, some of which are publicly available. Presently, published metrics predominantly centre on the performance of internal moderation processes, encompassing content removal rates, enforcement actions against accounts, appeals and restorations.

Public authorities worldwide increasingly recognize the importance of holding online platforms accountable for maintaining safety standards and protecting users' rights. For example, legislation such as the European Union's Digital Services Act (DSA)[8] and Digital Markets Act (DMA)[9] aim to establish clear rules for online platforms, including obligations to combat illegal content and ensure children's online safety. Similarly, the United Kingdom's Online Safety Act and Australia's Online Safety Act target online harms,[10] emphasizing the need for platforms to conduct thorough risk assessments and demonstrate that they have taken appropriate steps to protect their users.

Stakeholders in diverse sectors have long demanded the disclosure of metrics related to digital safety, particularly regarding platform transparency reporting. Transparency reporting – which was originally a voluntary practice before regulatory mandates were introduced – is becoming a cornerstone of efforts to promote awareness about digital safety. Major platforms have responded by publishing various metrics, often focusing on the efficacy of internal moderation processes and compliance with external requests for user data.



1.2 The challenge of measuring digital safety

Where metrics are developed and implemented to assess the effectiveness of interventions, they must be monitored over time to ensure they remain fit for purpose.

Despite these efforts, understanding and measuring digital safety outcomes remains complex. As previously described, this is due to a number of factors, including the dynamic nature of technological advances, the diverse array of digital products and services and the concurrent evolution of harmful behaviours. The picture is further complicated by the vast volume of digital content and the contextual or subjective nature of certain types of harm. Contextual factors such as culture and language are undeniably important, yet they are difficult to capture in a standardized or replicable way.

The challenge of crafting effective metrics to define and quantify phenomena that elude concrete and causal measurement was referred to as "measuring the immeasurable" by experts who contributed to the creation of this paper. Persistent tension exists between the benefits of comparable metrics across services or over time and the need for flexibility to account for differences in service characteristics, changing circumstances and the imperative for continuous adaptation and improvement. Additionally, stakeholders' varying ability to digest and use the information provided effectively further complicates the measurement of digital safety.

There is a delicate balance between the need for metrics to be accessible and understandable by various stakeholders (including users, policy-makers and investors) and their requirement to contain essential contextual information. Significant variations exist, even when consistent metrics are reported. For instance, "number of takedowns" may differ in granularity in terms of harms, confidence levels, documenting sources and reasons for takedown.

Several limiting factors compound these challenges, including: restrictions on sharing proprietary information or personally identifiable data; resource implications for measuring phenomena and tracking metrics; and the risk of reverse engineering or compromising internal systems and processes designed to increase user safety. Reporting on external requests for user data, such as legal demands and compliance rates, may also be restricted by gag orders or secrecy requirements.

Where metrics are developed and implemented to assess the effectiveness of interventions, they must be monitored over time to ensure they remain fit for purpose. One factor to consider is the potential for distorted incentives or unintended consequences when a particular metric becomes the basis for decision-making or performance targets. This can lead individuals or organizations to manipulate their behaviour to optimize that metric, resulting in a phenomenon known as Goodhart's Law:[11] "When a measure becomes a target, it ceases to be a good measure."

These challenges underscore the need for a comprehensive and nuanced approach to measuring digital safety that addresses the intricacies of the ever-evolving online environment. The next section aims to provide a categorization that can promote a more consistent approach to measuring digital safety outcomes.

2 Measuring digital safety

Proposed metrics for digital safety measure the impact on individuals, the risk of harm and safeguarding processes.

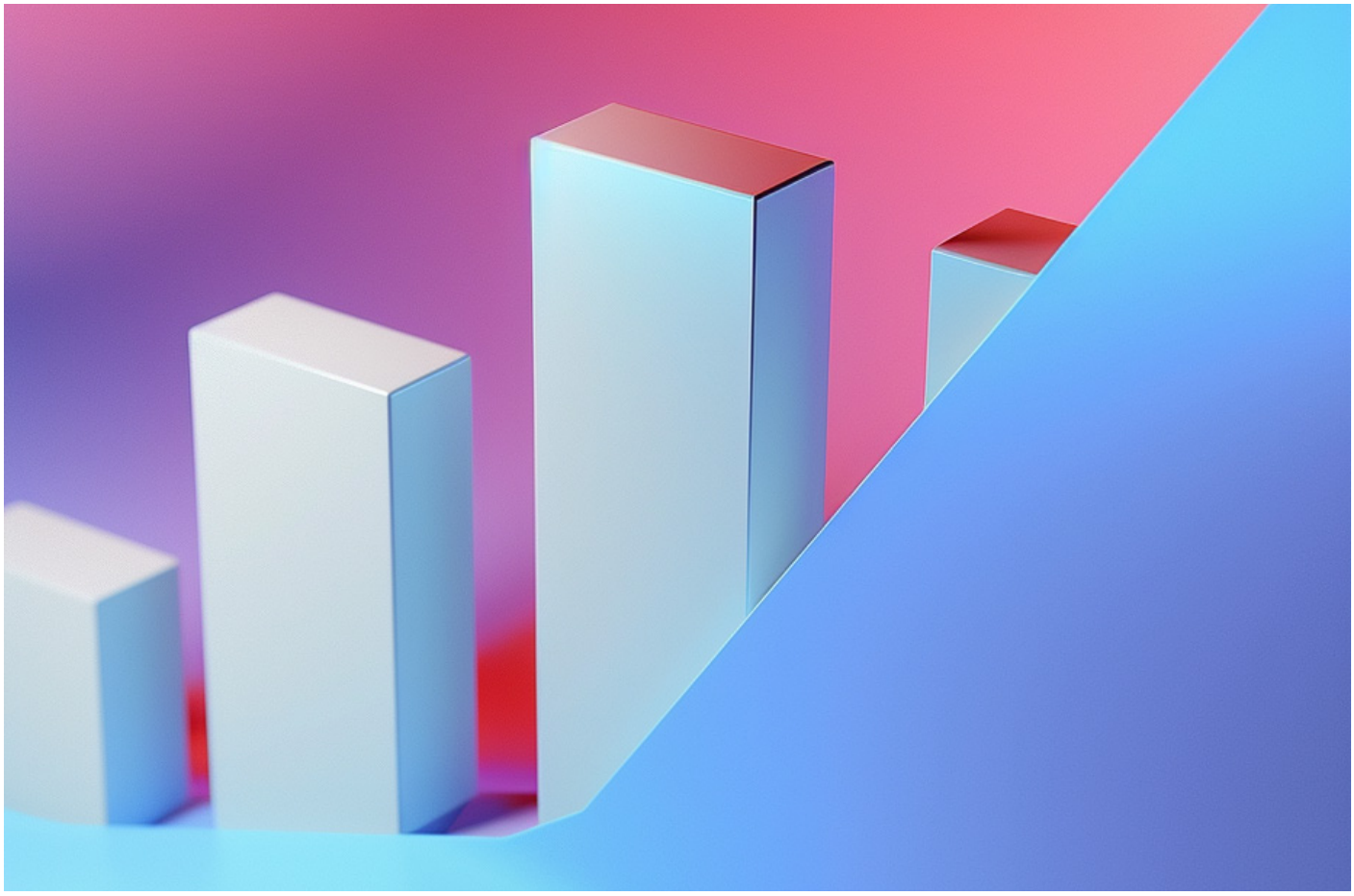
2.1 Categorization

Recognizing the complexity of measuring digital safety, this paper proposes grouping metrics into three categories:

- **Impact:** Metrics focused on translating subjective user experiences into quantifiable and objective data related to content or conduct, shedding light on user harms or benefits within the digital realm. This category may also cover the unintended impacts of interventions – for example, changes in how users behave or express themselves online. A comprehensive approach to measuring impact requires diverse data sources, including platform-specific data and insights from external stakeholders such as researchers and members of civil society.
- **Risk:** Metrics essential for identifying elements that increase the likelihood of users experiencing harmful outcomes on digital platforms. These metrics aim to improve prediction and prevention efforts in digital safety by measuring service features that elevate the risk of user harm. This includes metrics derived from various service-specific characteristics, such as user demographics and incentive structures driving user interaction and content distribution.
- **Process:** Metrics that assess the effectiveness of operational systems and processes implemented to mitigate digital harms and promote positive outcomes. These metrics provide indicators of overall intervention success. They evaluate operational outcomes throughout the life cycle of relevant systems and processes, emphasizing the need for transparent documentation from design and governance to execution and review.

Together, these metrics – explained in greater detail below – form a comprehensive framework for understanding and addressing digital harms, ensuring a safer and more resilient online environment for all users. By intertwining insights from impact, risk and process metrics, stakeholders can develop targeted interventions, enhance predictive capabilities and improve digital safety measures to safeguard users across digital realms. Each of the three categories has unique attributes that may serve as a guiding principle for organizing discussions and approaches related to digital safety measures, supplementing existing frameworks and enhancing the structured approach to digital safety assessment.

FIGURE 1 **Three types of metrics**



2.2 Impact metrics

Impact metrics seek to translate subjective user experiences into tangible, quantifiable data. They shed light on the outcomes and impacts of user experiences by measuring actual harm or positive impacts.

Proposals include:

1. **Understanding the experience of harm through collaboration with experts:** Collaborating with expert organizations focused on particular harms, including survivor advocacy groups, law enforcement agencies and independent research entities, can be valuable for understanding the impact of online harm, including for user groups that are disproportionately susceptible to particular harms.

2. **Uncovering patterns:** Analysis of data reveals patterns in behaviour. This can be invaluable for identifying cases of victims/offenders where harms have occurred or for helping to tease out causal relationships where interventions have been successful.

A comprehensive approach involving a wide range of data sources is essential for measuring impact effectively. This includes both platform-specific data and insights from external stakeholders such as researchers and members of civil society. However, platforms should refrain from collecting some off-platform data regarding the occurrence of harms (mainly for privacy reasons). It is therefore crucial for them to collaborate with specialist organizations to ascertain the most suitable approaches for collecting targeted data while avoiding privacy issues. In situations involving severe adverse effects and significant harm, engagement may extend to survivor advocacy organizations and law enforcement agencies.

User groups disproportionately susceptible to experiencing harm should be identified. Stakeholder insights are instrumental in tailoring interventions and heightening protective measures for these at-risk populations. This involves measuring the scale, severity and likelihood of harm occurring and identifying groups deeply affected by harm through insights into user experiences.

Examples of potential impact metrics and data sources are outlined in Table 1.

TABLE 1 | **Impact metrics and data sources**

Measurement	Impact metric	Description
User experience	Volume	Number of individuals affected: the quantification of individuals affected by the issue, providing a clear understanding of the scale of the problem. Information can be gathered from users reports/complaints.
	Severity	Impact on affected individuals (directly and indirectly): detailed examination of the consequences experienced by both directly affected individuals and those indirectly influenced by the issue, encompassing emotional, financial and societal ramifications. A possible data source could be front-line groups.
	Permanence	Duration or remediability of impact: assessment of the duration of the impact and the potential for remediation, highlighting whether the effects are temporary or persistent and if they can be mitigated or reversed. Reports from human rights organizations could be a data source.
Affected groups	Demographic breakdown	Characteristics of affected individuals/groups: identification of specific demographic, socioeconomic or behavioural characteristics shared by the affected individuals or groups, aiding in targeted interventions and support. Front-line groups could also be a data source.
	Safety perception	Users' perceived safety on platforms: evaluation of users' subjective feelings of safety while engaging with the platform, including factors such as trust, security measures and transparency in addressing safety concerns. Possible data sources are surveys, feedback mechanisms and sentiment analysis.
Patterns of behaviour	Harm archetypes	Identifying commonalities across lived experience of harms: recognition of common themes or patterns in the experiences of individuals who have encountered harm, facilitating the development of preventive measures and support initiatives. Users reports and complaints can be data sources.
	Offender data	Identifying commonalities across violative accounts: analysis of shared characteristics or behaviours exhibited by accounts involved in violating platform policies or causing harm, informing moderation strategies and content enforcement efforts. Possible data sources are repeat offenders, bots or malicious accounts identified and reports to child protection agencies, etc.
	Changes in user behaviour	Observable changes in usage patterns: examination of shifts in user behaviour or engagement with the platform that may indicate emerging or evolving issues, guiding proactive responses and adjustments to platform policies and features. Possible data sources include uptake of user empowerment tools and user churn rate, etc.

2.3 Risk metrics

Risk metrics measure issues with content, product, policies and processes and can inform the risk of user harm. While exposure to certain content or behaviour does not necessarily result in negative/positive outcomes, the overall prevalence of harmful material on a platform can influence the likelihood of users experiencing harm.

Factors such as the nature of the service, the user base and relevant harms must be considered. To understand the relevant harms, the Typology of Online Harms[12] can help to inform risk metrics by providing a list of harms and common terminology.

Identification, detection and prevention are key categories in risk metrics, highlighting the need for strategies that address both the content and the actors behind potential harms.

Examples of potential risk metrics and data sources are listed in Table 2:

TABLE 2 | **Risk metrics and data sources**

Measurement	Risk metric	Description
User base	Monthly active users	Number of accounts; humans interacting: quantification of active accounts and the number of individuals engaging on the platform, providing insights into the platform's user base and interaction levels.
	Demographic metrics	Children vs. adults: segregation of users based on age demographics, allowing for tailored interventions and policies to protect vulnerable groups such as children. Gender: analysis of user gender distribution to understand potential gender-specific issues and preferences in content consumption and interaction. Language used for each demographic: examination of the languages used by different demographic groups, aiding in multilingual content moderation and support initiatives.
	Human interaction metrics	Communication patterns and behavioural signals: exploration of user interaction patterns and behavioural cues to detect and address potentially harmful behaviours or content.
Prevalence	Percentage of identified content	Breakdown of identified content based on the method of detection: categorization of flagged content by the detection method employed, including proactive monitoring, human moderation, user reports and input from trusted sources, offering insights into the efficacy of moderation mechanisms.
	Accuracy of content moderation	Accuracy of human moderation against tested datasets: evaluation of the precision and reliability of human moderation in identifying and addressing harmful content, benchmarked against validated datasets.
	Method of identification	Effectiveness of different identification methods: assessment of the efficiency and accuracy of various content identification approaches, including algorithmic detection and input from trusted flaggers, to optimize moderation strategies.
	Content categorization	Classifiers to categorize harmful content: Implementation of classifiers to categorize harmful content types such as child sexual abuse material (CSAM), enhancing the platform's ability to swiftly detect and remove illicit content.
	Trusted flaggers by impact group	Identification of trusted individuals or groups contributing to content moderation efforts: recognition of reliable contributors to content moderation efforts and their impact on mitigating harmful content proliferation.
	Threat management	Speed of threat identification and resolution: measurement of the timeliness in identifying and addressing emerging threats, ensuring swift responses to safeguard user safety and platform integrity. Integration of new threats into business as usual (BAU): evaluation of the normalization rate of newly identified threats within standard business operations, guiding adjustments to response protocols and preventive measures.
	Testing and optimization	Testing metrics for new products and features: Use of system-readiness level frameworks to assess the readiness and performance of new platform features and products, minimizing potential risks to user safety.
Incentive	User engagement	Measurement of user engagement incentives: evaluation of user engagement metrics such as dwell time, click-through rates (CTR) and churn rate to understand their influence on user exposure to potentially harmful content.
	Harmful content views	Measuring how potentially harmful content is viewed: analysis of how users encounter potentially harmful content, including exposure via algorithms, recommendations or user-generated sharing, to refine content distribution algorithms and moderation strategies.
	Alignment with regulatory compliance	Assessment of platform incentive structures: evaluation of how platform incentive structures prioritize compliance with regulations, human rights standards and community safety, ensuring alignment with overarching safety objectives.

Note: The data source for each entry is company data.

2.4 Process metrics

Process metrics evaluate the effectiveness of the tools and processes implemented to mitigate digital harms, serving as indicators of the overall maturity of such interventions.[13] They measure operational outcomes from systems and processes intended to improve digital safety, spanning design, implementation and continuous improvement of systems relating to digital safety. Process metrics can document digital safety practices across the product development life cycle, from the initial design and governance stages to the enforcement and improvement phases, emphasizing the necessity of transparent and rigorous documentation.

Effectiveness is best evaluated through independent external evaluation (e.g. such as a third party assessor akin to the Digital Trust and Safety Partnership model or audits akin to formal regulation) during which systems and processes for ensuring risk mitigation can be tested.

Examples of potential process metrics and data sources are provided in Table 3.

TABLE 3 **Process metrics and data sources**

Measurement	Process metric	Description
Product/policy development	Error rate	Organization's ability to address errors: assessment of the organization's capability to respond to failures or issues without endangering user safety.
	Number of hits to relevant pages	Frequency of user access to specific pages such as terms of service, privacy policies or transparency reports: exploration of how often users visit important informational pages, indicating their engagement with the platform's policies and transparency efforts.
Enforcement	Accuracy of identification, whether using automated technologies or human moderators	Effectiveness of system or human moderator in identifying and categorizing items: evaluation of how accurately moderators classify various types of content, such as images, text sentiment, anomalies in data or individuals in images.
	(Successful) appeals	Number or percentage of successful appeals made by users or organizations: tracking of instances of users challenging actions taken by the platform and which result in a favourable outcome for the user. Successfulness of appeals processes also includes giving users channels for feedback.
	Response time	Speed and efficiency of platform reaction to safety issues or security threats: measurement of how quickly and effectively the platform responds to potential safety issues, security threats or incidents reported by users.
Improvement	Baseline comparisons when introducing new systems or processes	Evaluation of metrics in relation to established baselines or industry standards: comparison of the organization's safety and security metrics against established benchmarks or industry standards to assess performance.
	Trusted flagger analysis/impact	Identification and collaboration with trusted flaggers: evaluation of how effectively the organization works with trusted partners or entities to enhance threat detection and management.
	Integration into business as usual (BAU)	Integration time for new threats into standard business operations: assessment of the speed at which the organization incorporates newly identified threats into its standard operating procedures to mitigate risks effectively.

Note: The data source for each entry is company data.

3 Practical application of digital safety metrics

Metrics play a vital role in holding digital services accountable to user communities and regulatory requirements.

The practical application of digital safety metrics is critical for assessing current safety measures and guiding future improvements.

3.1 Drawing in other metric frameworks

Ultimately, metrics and measurements for online safety should be tailored to the specific goals and challenges of the digital environment.

When evaluating online safety, mature frameworks – such as those used in cybersecurity, environmental, social and governance (ESG) and the United Nations Guiding Principles on Business and Human Rights metrics[14] – can provide valuable guidance on defining and measuring effectiveness.

These frameworks provide established methods and indicators for assessing performance and risk within complex systems, offering insights that can be adapted to the context of online safety.

In cybersecurity, metrics often centre on key performance indicators (KPIs) related to threat detection, threat prevention, incident response times and system resilience. Similarly, within online safety, metrics could encompass measures such as platform responsiveness to reports of harmful content, the efficacy of content moderation algorithms in identifying and removing toxic and illegal material and the speed of addressing security vulnerabilities that could compromise user safety.

ESG metrics present another valuable perspective for evaluating online safety. Within ESG investing, companies are assessed based on their environmental impact, social responsibility and corporate governance practices. Similarly, online platforms could be evaluated based on their efforts to promote a safe and inclusive online environment, and the transparency of content moderation policies. 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.

The World Economic Forum's *Measuring Digital Trust* white paper is a comprehensive model for assessing digital trust within organizations.[15] By evaluating organizational maturity across various dimensions of digital trust, the paper aims to gauge the effectiveness of governance structures in meeting individual and organizational expectations. This evaluation aligns with the dimensions outlined in the digital trust framework, covering decision-making, cybersecurity, safety, transparency, interoperability, auditability, redressability, fairness and privacy. Unlike measures tracking progress towards overarching digital trust goals, these metrics focus on examining the internal objectives and capabilities of an organization's digital trust programme. The measures are characterized by their objectivity and forward-looking nature, providing valuable insights into an organization's readiness to navigate the complexities of digital trust in an ever-evolving digital landscape.

Understanding the implications of online safety metrics requires careful consideration of context and nuance. For instance, a decrease in reported incidents of online harassment may suggest improved platform safety measures, but it could also indicate underreporting due to user distrust in the reporting process. Similarly, an increase in the speed of content removals may reflect proactive moderation efforts, but it could also hint at overzealous censorship that stifles free expression.

Ultimately, metrics and measurements for online safety should be tailored to the specific goals and challenges of the digital environment. Drawing on insights from established frameworks – such as cybersecurity, ESG and the Digital Trust Metrics Framework[16] – while remaining flexible and adaptable to evolving threats and user needs can help stakeholders develop a more comprehensive understanding of online safety and strive for continuous improvement in this critical domain.

3.2 Addressing access to data

Stakeholders must, by recognizing the critical need for a diverse dataset in assessing digital safety, collaboratively determine which entities are best positioned to access this data. This involves striking a delicate balance between the imperative for comprehensive analysis and privacy and security considerations. As the risk framework highlights, it is important to ask questions about how personal data is managed in terms of storage and how it is shared with third parties.

Ongoing projects that take a multistakeholder approach provide valuable blueprints for enhancing data accessibility. For instance, initiatives involving collaborations among tech companies, academic institutions, government agencies and civil society organizations – such as the Global Coalition for Digital Safety – serve as models for using diverse expertise and perspectives to improve the quality and utility of digital safety metrics. These collaborative efforts not only facilitate the sharing of data but also promote interdisciplinary dialogue and innovation in addressing digital safety challenges.

Moreover, there is an opportunity for further collaboration, particularly regarding researcher access to data. Many researchers face significant barriers when accessing relevant data to study digital safety trends and phenomena. Streamlining processes for data access and promoting partnerships between researchers and data custodians in a privacy-protecting way can enhance data availability for research purposes, leading to more robust and evidence-based approaches to measuring and addressing digital safety issues.

3.3 Continuously improving practices and increasing accountability

It is essential to continuously improve practices and increase accountability in digital safety to create a safer online environment. The practical application of digital safety metrics plays a crucial role in achieving these goals by allowing for the critical evaluation of interventions and gauging their effectiveness in real time.

Insights gleaned from best practices across various platforms inform the continual development and refinement of design principles and operational strategies, contributing to iterative progress in digital safety and user protection. By analysing the outcomes of different safety measures, organizations can identify areas for improvement and implement targeted interventions to address emerging threats and shifting challenges.

> By analysing the outcomes of different safety measures, organizations can identify areas for improvement and implement targeted interventions to address emerging threats and shifting challenges.

Digital safety metrics reinforce accountability within the online environment. By offering both quantitative and qualitative data, these metrics empower platforms to enact meaningful change to improve digital safety while providing NGOs and regulatory bodies with the insights they require to effectively oversee and evaluate the safety measures implemented by digital service providers. This comprehensive approach enables stakeholders to gain deeper insights into the effectiveness of various safety initiatives, allowing for informed decision-making and targeted interventions where necessary.

Moreover, such metrics serve as essential benchmarks for monitoring compliance with applicable legislation and regulatory frameworks, thereby promoting a safer online environment for users. Digital safety metrics also enhance user trust and confidence in online platforms by ensuring that platforms adhere to or exceed the compliance standards intended to promote a safer online environment. Metrics are only one part of safety assessment, though; evaluation should be based not on how many parameters are being reported on, but on what makes the most sense for a service provider to report, bearing in mind the nature of service and the type of harm being addressed.

By incorporating impact, risk and process metrics into their approaches to risk management and external reporting, stakeholders can build a better understanding of digital safety issues and develop targeted interventions to tackle specific challenges.

This structured approach not only enhances accountability but also promotes continuous improvement in digital safety practices, ultimately leading to a safer and more secure online environment for all users. However, it is important to emphasize that regulatory requirements should be harmonized globally to make measurement truly effective.

4 Emerging considerations

In looking to the future, risk assessment and quality-checking of interventions are vital, along with a commitment to increased transparency.

To understand the full spectrum of harms and their impacts, it is essential to prioritize risk assessment. Effectively identifying and mitigating risks can help reduce the likelihood of unintended consequences from the use of platforms and services. The Global Coalition for Digital Safety has already produced the *Digital Safety Risks Assessment Framework in Action*,[17] and this paper aims to improve the robustness of such assessments by providing fresh, actionable insights into metrics informed by expert multistakeholder discussions.

It is crucial to assess interventions to determine whether safety measures are having their intended effects and demonstrating their impact on user safety. This work will be complemented by the next output of the coalition: a report on effective interventions that online service providers are implementing to mitigate online harms.

A commitment to continuous improvement is vital. Platforms must continuously evolve to enhance user safety, using insights gleaned from examining best practices and from lessons learned from their interventions and approaches to safety to inform platform design and regulatory frameworks.

Increased accountability is paramount. In this sense, metrics and measurements are essential to help increase transparency. They can support NGOs and regulators in holding platforms accountable, and ensure compliance with relevant legislation, which is essential for promoting a safer digital environment.

Conclusion

Metrics that measure online safety promote targeted, effective governance to create a more secure and less threatening online environment.

This white paper, which is intended to be a comprehensive guide for understanding safety metrics and measurements in the realm of digital services, offers insights for platforms, regulators and users as they navigate the complex landscape of online harms. While it does not provide an overview of all metrics and measurements, the paper aims to provide a method of categorization to tackle the challenge of what is termed “measuring the immeasurable”.

Establishing metrics for online safety is crucial for good governance as it promotes accountability, aids evidence-based decision-making, monitors progress, guides resource allocation, facilitates benchmarking, promotes transparency and engagement and enables the evaluation of intervention effectiveness.

Measuring online safety is imperative for cultivating a safer and more resilient digital environment. It enables informed decision-making, facilitates policy development and enhances stakeholders’ awareness of online safety issues. By embracing these approaches, stakeholders can collectively work towards a safer online environment for all.

Contributors

Lead Authors

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

Collin Kurre
Technology Policy Principal, United Kingdom Office of Communications (Ofcom)

David Sullivan
Executive Director, Digital Trust and Safety Partnership.

World Economic Forum

Minos Bantourakis
Head of Media, Entertainment and Sport Industry

Daniel Dobrygowski
Head, Governance and Trust

Daegan Kingery
Specialist, Digital Safety and Trustworthy Technology

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

Acknowledgements

Maria Cristina Capelo
Head of Safety Policy, Meta Platforms

Jeffrey Collins
Director, Trust and Safety, Amazon Web Services

Julie Dawson
Chief Policy & Regulatory Officer, Yoti

Julia Fossi
Trust & Safety: Global External Engagements Lead, Amazon Web Services

Susie Hargreaves
Chief Executive Officer, Internet Watch Foundation

Sasha Havlicek
Chief Executive Officer, Institute for Strategic Dialogue

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

Adam Hildreth
Founder, Crisp Thinking Group

Julie Inman Grant
Commissioner, Australia eSafety Commissioner

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

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

Angela McKay
Director, T&S Research & Partnerships, Google

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

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

Bo Viktor Nylund
Director of UNICEF Innocenti Global Office of Research and Foresight, United Nations Children's Fund (UNICEF)

María Paz Canales Loebel
Head of Legal, Policy and Research, Global Partners Digital

Katherine Sandell
Head of Scale and Convergence, Core Risk Office, Google

Noam Schwartz
Chief Executive Officer and Co-Founder, ActiveFence

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

Liz Thomas
Director Public Policy, Digital Safety, Microsoft Corporation

Sam Wallace
Research Analyst, Sustainability and Lead, Technology and Communications Sector, International Financial Reporting Standards Foundation (IFRS Foundation)

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

David Wright
Chief Executive Officer, SWGfL

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

Production

Laurence Denmark
Creative Director, Studio Miko

Alison Moore
Editor, Astra Content

Oliver Turner
Designer, Studio Miko

Endnotes

1. World Economic Forum. (2023). *Digital Safety Risk Assessment Framework in Action*. https://www.weforum.org/publications/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies/.
2. Ibid.
3. World Economic Forum. (2023). *Global Principles on Digital Safety: Translating International Human Rights for the Digital Context.* https://www3.weforum.org/docs/WEF_Global_Charter_of_Principles_for_Digital_Safety_2023.pdf.
4. World Economic Forum. (2023). *Digital Safety Risk Assessment Framework in Action.* https://www.weforum.org/publications/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies/.
5. World Economic Forum. (2023). *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.
6. World Economic Forum. (2023). *Global Principles on Digital Safety: Translating International Human Rights for the Digital Context.* https://www3.weforum.org/docs/WEF_Global_Charter_of_Principles_for_Digital_Safety_2023.pdf.
7. Google. (2021). "Google Transparency Report". https://transparencyreport.google.com/about.
8. European Commission. "The Digital Service Act". https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.
9. European Commission. "The Digital Markets Act". https://digital-markets-act.ec.europa.eu/index_en.
10. United Kingdom Parliament. (2023). "Online Safety Act 2023". https://bills.parliament.uk/bills/3137; Australian Government e-Safety Commissioner, "Learn About the Online Safety Act". https://www.esafety.gov.au/newsroom/whats-on/online-safety-act.
11. Oxford Reference. (2024). "Goodhart's Law". https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095859655.
12. World Economic Forum. (2023). *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*. https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.
13. Digital Trust & Safety Partnership. *Digital Trust & Safety Partnership Best Practices Framework*. https://dtspartnership.org/best-practices/.
14. United Nations. (2011). *Guiding Principles on Ethics and Human Rights*. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.
15. World Economic Forum. (2023). *Measuring Digital Trust: Supporting Decision-Making for Trustworthy Technologies*. https://www.weforum.org/publications/measuring-digital-trust-supporting-decision-making-for-trustworthy-technologies/.
16. Ibid.
17. World Economic Forum. (2023). *Digital Safety Risk Assessment Framework in Action*. https://www.weforum.org/publications/digital-safety-risk-assessment-in-action-a-framework-and-bank-of-case-studies/.

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