WORLD ECONOMIC FORUM

Preparing for the Future: The Pillars of Digital Safety Foresight

INSIGHT REPORT

JUNE 2026

Images: Getty Images

Contents

Foreword

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

David Sullivan
Executive Director, Digital Trust and Safety Partnership

Agustina Callegari
Initiatives Lead, Technology Governance, Safety and International Cooperation, World Economic Forum

Cathy Li
Head, Centre for AI Excellence; Member of the Executive Committee, World Economic Forum

In a rapidly shifting digital environment, regulators, civil society, technology companies and others are all grappling with the need not only to respond to today's risks but also to anticipate and prepare to address future harms.

The purpose of this insight report is to provide a clear and practical approach to digital safety foresight work. Organizations already conduct forms of foresight. This report seeks to help move such work beyond abstract scanning or speculative discussion to the capacity to anticipate, interpret and act on change in specific contexts. Foresight, in this context, is not about predicting the future with certainty, but strengthening readiness and creating the structures, signals and decision processes that enable institutions to respond more coherently. This supports a safety-by-design approach that invests in risk mitigation at the front end, supporting organizations to embed user protections from the get-go and avoid costly retrofits.

The digital safety foresight pillars presented in this report – 1) Frame and govern, 2) Sense, 3) Model, 4) Assess readiness, 5) Install early warning and 6) Synthesize – structure best practices for organizations to conduct foresight. These pillars were developed through extensive consultation with practitioners and experts across the digital safety ecosystem. The report builds on those pillars and outlines four response pathways to help organizations determine the next steps based on foresight outputs – such as tightening existing protections, testing and learning, building new safeguards and establishing tripwires where uncertainty remains high.

The use of foresight in digital safety plays a key part in stopping harms from growing and spreading. As harms can move across services, sectors and borders, collaboration is essential. This report continues to advance the Global Coalition for Digital Safety's goal of strengthening organizations' preparedness and effectiveness in digital safety work, while contributing to a more coordinated, resilient and anticipatory digital safety ecosystem.

Executive summary

Combining six pillars and four response pathways, this report provides a practical guide to foresight in organizations.

Preparing for shifts in the digital environment has always been critical for organizations, but it is especially urgent now as new technologies – such as agentic AI, genAI and the convergence of wearables and immersive platforms (VR/AR) – create new vectors for harm that adversaries are quick to adopt and exploit. Foresight helps business teams, particularly trust and safety personnel, to anticipate emerging harms before they scale and to plan responses with the expectation that some harms will still succeed. Organizations already use foresight in this way, but some, especially those with limited capacity, lack a dedicated process or mechanism. This foresight report builds on the Global Coalition for Digital Safety's 2023 report *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms* and the 2025 report *The Intervention Journey: A Roadmap to Effective Digital Safety Measures*. The foresight pillars presented here distil best practices from consulted organizations into six core elements of digital safety foresight:

Frame and govern: Sets the question, boundaries and decision rights.

Sense: Expands an organization's field of view beyond what is already measured, already debated internally or already visible through day-to-day operations.

Model: Turns signals of change into decision-relevant explanations of how harms could emerge, scale and adapt in the real world.

Assess readiness: Tests whether an organization can credibly prevent, detect, respond to and learn from the kinds of digital safety harms it may face.

Install early warning: Translates modelling into ongoing awareness and support efforts in ensuring readiness.

Synthesize: Consolidates outputs from the other pillars into a single, decision-ready view of an emerging harm.

The pillars are not intended to be a prescriptive set of steps, but rather a guide to the elements most important for effective digital safety foresight. Based on the outputs of the pillars, the report also presents four response pathways to help orient organizations on next steps. The four paths are:

1. **Known and covered:** Where organizations tighten existing protections.
2. **Novel/uncertain coverage:** Where teams test, learn and decide whether the organization's posture must change.
3. **Not covered:** Where new safeguards should be built before exposure scales.
4. **Unpredictable/unknown:** Where tripwires, sentinels and strong redress help organizations act cautiously under uncertainty.

Through the digital safety foresight pillars and response pathways, this report provides a foundational structured approach for foresight within organizations and ensures such work moves beyond a knowledge and speculative activity to a more action-oriented process.

Introduction

To combat potential threats in a fast-developing digital world, organizations will benefit from sharing best practices.

> It is not enough to act only after a harm occurs and scales; organizations need to spot signs of change that can result in new vectors and harms.

Today's online world is changing constantly. Rapid shifts in AI, synthetic media, private messaging services and immersive digital environments are reshaping how people influence, how trust is fostered and how quickly harmful behaviour can spread. These shifts widen the opportunities for harms, such as fraud, coercion, harassment, grooming, abuse-of-service and privacy-linked exploitation. The harms may arise through visible content and interpersonal interaction but also through the misuse of enterprise, cloud and infrastructure services that are exploited to harm third parties. It is not enough to act only after a harm occurs and scales; organizations need to spot signs of change that can result in new vectors and harms.

A defining feature of this landscape is that adversaries are learning and adapting across ecosystems. Harm is increasingly distributed among organizations, while low-cost automation and synthetic identity reduce the friction required to probe controls, impersonate trusted actors, bypass protections and manipulate victims.

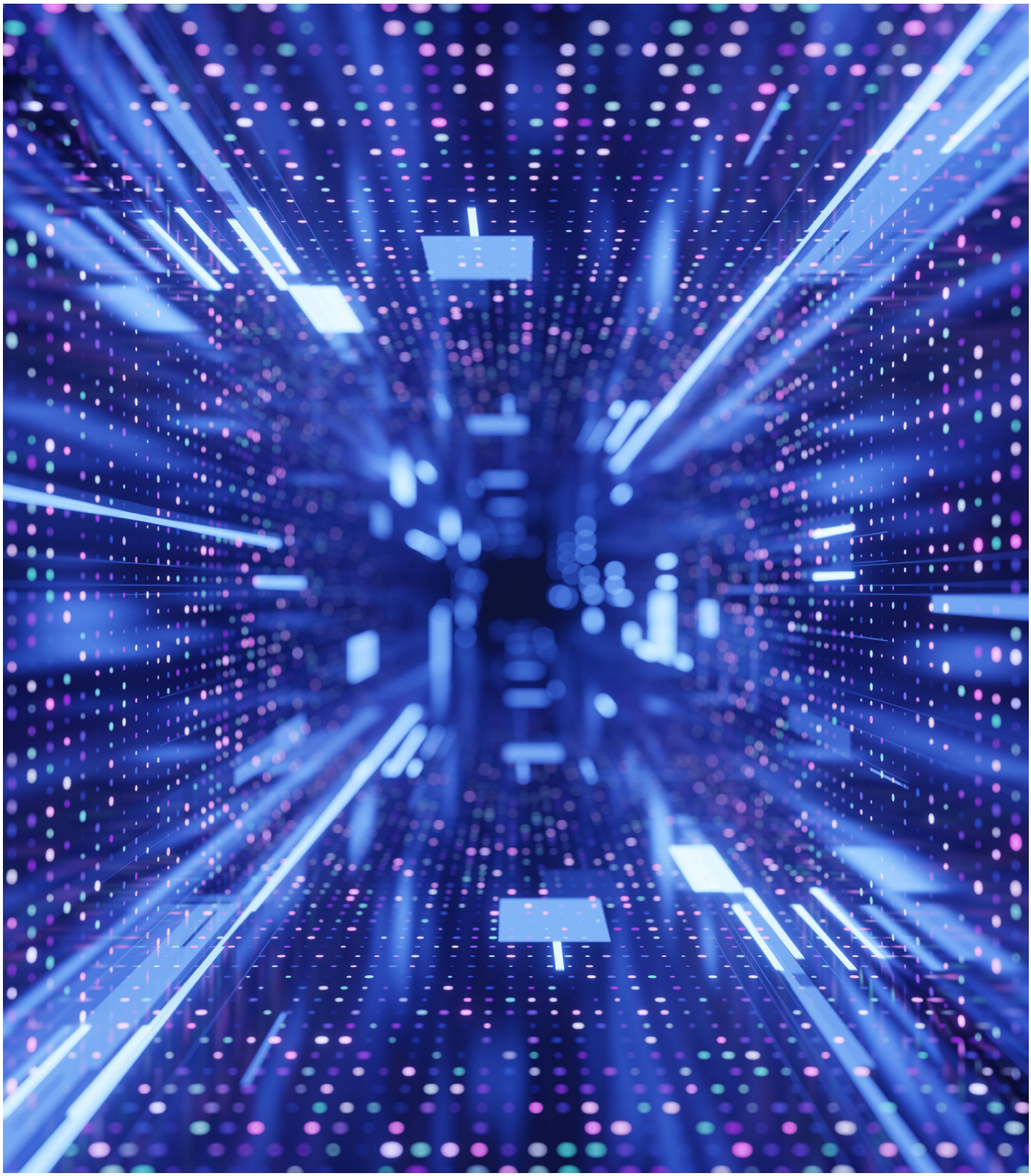
Meanwhile, the governance environment is evolving alongside the threat landscape, but not always evenly. Systemic-risk regimes, online safety requirements and privacy laws are pushing organizations towards ongoing risk management and accountability, while AI governance, privacy regulation and sector-specific rules introduce overlapping duties and changing expectations. Operational capacity is uneven: some organizations have strong and mature trust and safety, threat intelligence and readiness functions, while others operate with small teams, have limited visibility and may deprioritize foresight tasks. The result is an uneven "safety floor", where similar harms can look very different depending on where evidence exists, who has authority to act and what tools an organization has.

Many existing approaches already address safety risks effectively, particularly those tied to internal product changes, such as new features, policy updates or model releases. That inward-looking perspective remains essential. However, organizations also need to look externally at changes outside their own systems: shifts in technology, regulation, global events or attacker tactics can create new risks.

The goal of this report is not to introduce foresight as a novel concept, but to recognize and strengthen what many teams are already doing by sharing best practices from organizations in the form of six digital safety foresight pillars. It also offers four response pathways to help provide general directions for organizations based on the results of their foresight procedure. The report does not seek to prescribe a single correct workflow or assume large-team capacity, but it does seek to reinforce the need for foresight work in organizations. The pillars are designed to be usable and adaptable according to capacity, including suggesting ways to use AI to reduce workload while maintaining human accountability.

1 Drivers of change in the digital environment

Five fast-moving areas of development, including AI and neurotechnology, pose challenges for those seeking to promote digital safety.



The digital environment is changing at an accelerating pace as new AI capabilities mature, distribution channels fragment and recombine, business incentives shift and regulatory expectations evolve. The result is not only more incidents but faster-moving and more adaptive harm pathways – where tactics, targets and surfaces can change in a matter of weeks. This makes foresight increasingly important, as it helps teams identify emerging conditions early, challenge assumptions before they harden into blind spots and prepare proportionate responses in advance of harms scaling beyond manageable interventions.

The drivers below do not constitute a comprehensive catalogue; they highlight prominent external shifts that are reshaping how harm is generated, scaled, exploited and concealed across the wider digital ecosystem. The issue is not that these technologies are inherently harmful but that the same capabilities creating value for users and organizations also alter the safety surface in ways that foresight must anticipate. These drivers are already being addressed, but the challenge is that the underlying conditions are accelerating and recombining in novel ways, creating new vectors and spillover effects that may not match historic patterns or organizational assumptions.

1.1 AI and automation

Agents can amplify harassment and information campaigns by automating high-volume abusive interactions and coordination.

AI is shifting from being solely the discrete functionality of a chatbot to becoming a potent general-purpose layer that mediates communication, search, creation and decision-making through agentic AI. Automation is also becoming more anthropomorphic, with models now able to generate convincing language and media, sustain long conversations and, through tool use, take actions across digital environments. These capabilities benefit users and organizations, but they also reduce the cost and ease of influencing users, committing fraud and abuse, enabling adversaries to operate with greater speed, scale and plausibility than earlier "bot" eras.

Parasocial/AI relationships

Conversational systems are increasingly used as social partners, from companions, coaches and role-play characters to always-available confidants. For many users, these systems can provide companionship, study assistance, emotional support, accessibility and low-friction help at times when human support is unavailable. Survey work suggests that this is already mainstream among teenagers. Common Sense Media reports that a majority of US teenagers have tried AI companions,[1] and Pew Research Center finds widespread chatbot use among teenagers, with many reporting daily interaction.[2] The safety shift is not only that people interact with AI but that they form patterns of trust and disclosure with systems that can mirror tone, remember details and adapt to emotional cues. It is not just AI; online personalities are increasingly fostering (intentionally or not) parasocial relationships by sharing frequent personal updates, speaking directly to audiences and building community. The result can be bonds that feel friendship-like through perceived intimacy and authenticity.[3]

These relationship dynamics create new vectors for manipulation and grooming. Adversaries can deploy synthetic companions or conversational personas to cultivate rapport at scale, learning personal details over time, testing boundaries and nudging behaviour incrementally. Because interaction is private and persistent, manipulation can look like care, with prompts that normalize risky behaviour, encourage secrecy, redirect users off-platform or harvest sensitive disclosures for later coercion. Harms often appear as gradual dependency or escalation rather than a single policy-violating message, making detection, reporting and redress more difficult.

AI agents

A major technological shift is the rise of AI systems that can plan and execute tasks rather than simply responding to prompts. Browser and computer-use agents can navigate interfaces, click buttons and complete multistep workflows in environments built for humans.[4] This affects digital safety because it lowers the barrier for automation to participate wherever humans participate in supporting channels, marketplaces, onboarding flows, developer portals and internal enterprise tools without needing privileged application programming interface (API) access. In legitimate use, these agents can reduce operational burden, improve accessibility and help users and workers complete complex digital tasks more efficiently.

For adversaries, agentic automation expands both its scale and attack surface area. Agents can register accounts, probe friction points, generate convincing support tickets or run social engineering campaigns that adapt based on live responses. User interface (UI) agents can dramatically scale inauthentic participation that erodes trust in what appears popular and legitimate, an effect long associated with social bots, but potentially easier to execute across many services with agents. Agents can amplify harassment and information campaigns by automating high-volume abusive interactions and coordination.[5]

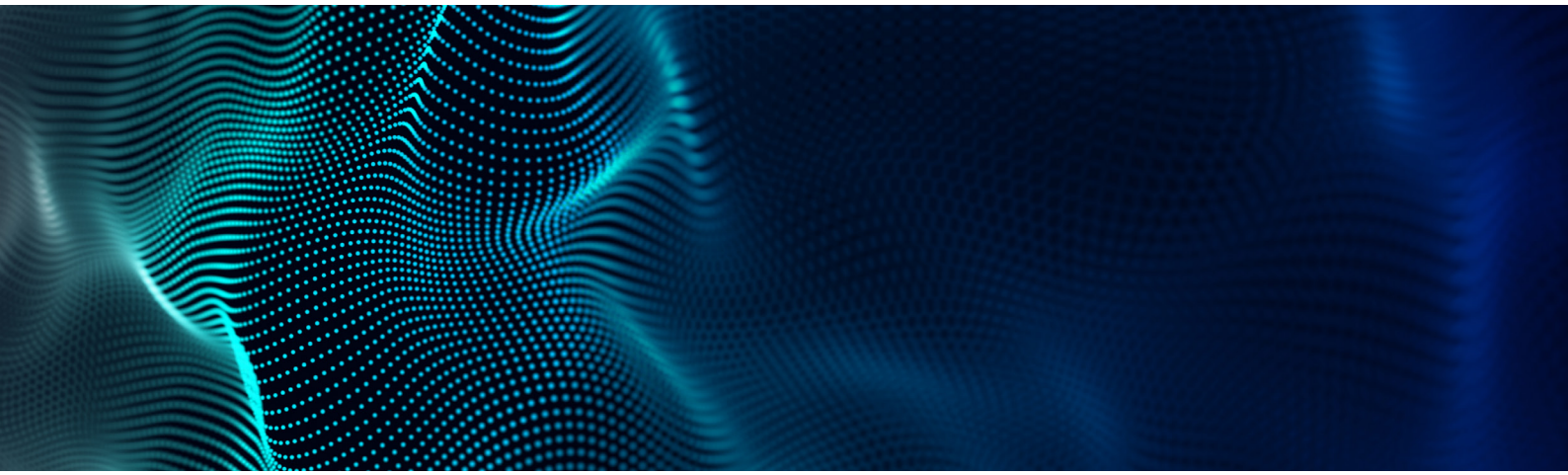
Agentic automation also accelerates infrastructure-level abuse, as agents can programmatically provision cloud resources, register domains, deploy phishing or malware-delivery infrastructure and cycle through accounts and IP addresses faster than manual

abuse operations.[6] They also introduce new attack classes: when agents consume untrusted content, they can be manipulated through prompt injection and related techniques, redirecting behaviour or extracting data. It has been noted that prompt injection is a persistent risk for web-based agents operating on open content.[7] This shifts some safety work from content moderation towards execution controls, such as least-privilege tool access, audit logs, transaction confirmation and the ability to halt or roll back actions when behaviour deviates.

AI and the information layer

Foundation models and AI assistants are becoming an interface layer for information as they summarize, recommend and answer questions in a single conversational flow. This can create real value by lowering search friction, speeding synthesis and helping users navigate large or complex information environments. Public adoption is rising quickly, including use for news and current affairs.[8] The risk is not only hallucination, but authority: when an assistant presents a confident synthesis, users may be less likely to inspect sources, and errors can propagate through reposts, screenshots and downstream decisions.

This changes the dynamics because persuasion becomes cheaper and more targeted. Attackers can generate tailored narratives, test variants for maximum engagement and flood information spaces with evidence-dense arguments that are persuasive but unreliable. In organizational settings, assistants also create new data flows, such as prompts and interaction logs that can include sensitive information, and adversaries may deliberately bait systems into revealing confidential information or generating harmful guidance. Safety teams therefore need to track how AI systems change both user behaviour and the organization's attack surface across products and workflows.



1.2 Synthetic media and identity

Synthetic media has moved from specialized tooling to everyday workflows. Voice cloning, AI dubbing, text-to-image/video and real-time filters now appear within consumer apps and enterprise products. These tools support legitimate use cases, such as localization, accessibility, creative production and identity expression. However, as these tools become cheaper and easier to use, the defining shift for safety is the erosion of default trust: visual and auditory cues that once carried strong evidentiary weight now require verification and context.

Synthetic information and identity

Synthetic information enables more believable fraud, coercion and impersonation. Financial regulators warn that deepfake audio and video are being used in fraud schemes targeting institutions and customers.[9] In one widely reported case, a firm was tricked into transferring roughly $25 million after fraudsters used deepfake video and audio in a conference setting.[10] These incidents illustrate a broader pattern of adversaries combining AI-generated media with compromised data, social engineering and rapid payment rails to exploit moments when verification is weakest.

Harms also extend to harassment and abuse. Research and industry reporting consistently find that a very high share of deepfakes online are non-consensual sexual content, with over 90% of deepfakes being non-consensual pornography and predominantly targeting women and girls.[11] Synthetic intimate images can be used for humiliation, extortion and coercion, and deniability can be weaponized both by perpetrators and by those seeking to dismiss legitimate evidence.

Identity itself is becoming more synthetic. GenAI can produce convincing profiles, documents, voice samples and interaction histories, enabling synthetic identity fraud that blends real and fabricated data to bypass verification.[12] For digital services, synthetic identities can open accounts, evade bans, infiltrate communities, abuse marketplaces and gain access to enterprise workflows that assume a human counterparty. Over time, synthetic content saturation can also degrade safety tooling. Research on "model collapse" shows that repeated AI model training on generated data can erode a model's performance and edge-case coverage first – precisely the examples needed to detect novel abuse and minority harms.[13]

Content provenance and authenticity

Provenance standards are emerging to preserve a record of the context in which media was created and modified. The Coalition for Content Provenance and Authenticity (C2PA) specification defines a way to attach cryptographically signed content credentials to images, video and audio.[14] Major platforms and tool providers have begun experimenting with these signals, including commitments to attach or display credentials for certain AI-generated content.[15] Where implemented well and widely, provenance can support faster triage, clearer user context and stronger accountability across reposts and syndication.

However, adoption is uneven and adversaries can route around credentials by stripping metadata, re-recording content or circulating through channels that do not preserve signatures, watermarks or metadata; they can also attempt spoofing through compromised keys or fake verified overlays.[16] Effective use therefore depends on end-to-end preservation across capture, editing, hosting and viewing plus UI that is resilient to manipulation and clear about what a credential does and does not guarantee.

1.3 Immersive and always-on sensing

Digital interaction is becoming more continuous and sensor-rich. Smartphones, wearables and ambient assistants make it easier to communicate and transact in real time, while immersive interfaces add spatial presence and multimodal cues. These shifts expand the kinds of data collected, the contexts in which harm can occur and the speed with which adversaries can reach targets.

Always connected and real time

Always-on connectivity, which is now a baseline condition of digital life, changes the tempo of harm, with 41% of US adults reporting being online "almost constantly" and higher intensity among younger adults.[17] Constant reach means less cool-down time, potentially making users less alert and more vulnerable to multiple forms of social engineering. Moreover, scams are escalating quickly through notification loops, urgent calls and multichannel pressure, while enterprise collaboration tools can be exploited for high-speed credential phishing and approval fraud.

Sensor-rich devices can be exploited, with location, contacts and behavioural traces combined with synthetic media to create tailored impersonation and coercion attempts. As more services integrate AI, the stakes increase to minimize sensitive data exposure.

Neurotechnology

Neurotechnology is moving towards consumer and workplace ecosystems through wearables, neurofeedback apps and emerging brain–computer interfaces. These tools may support rehabilitation, accessibility, well-being and new forms of human–computer interaction. The United Nations Educational, Scientific and Cultural Organization (UNESCO) adopted a Recommendation on the Ethics of Neurotechnology in 2025, reflecting concern that neural data can reveal highly sensitive information about cognition, emotion and health.[18] The shift is that one's inner life becomes data, increasing the risks beyond privacy into autonomy and freedom of thought.

In high-pressure environments, neuro data could enable coercive monitoring and discriminatory decisions based on inferred traits; in consumer contexts, neural or affective signals could be used to optimize persuasion. Even where organizations do not deploy neurotech directly, they may rely on vendors and device ecosystems that do, creating third-party dependencies and new categories of data exposure.[19]

Immersive interfaces

Immersive interfaces (VR/AR/mixed reality) are increasingly appearing in daily life, from living rooms for gaming and fitness, through to workplaces for collaboration and training, to camera-first apps at concerts and in shopping aisles.[20] Used well, they can improve training, collaboration, accessibility, entertainment and context-rich communication. However, these interfaces change the harm surface because interaction is embodied and immediate. Spatial audio, gaze and hand tracking/haptics can make harassment feel physically close, while synthetic avatars and voice filters enable real-time identity deception. AR also moves digital risk into physical settings where overlays can mislead users about what they see, persistent anchors can enable place-based harassment, and always-on cameras raise bystander privacy concerns.

The net effect is that harms may be harder to observe and reproduce. Evidence is often ephemeral – as with a live session or a fleeting overlay – and reporting may require new forms of capture, including replay snippets, spatial context or device telemetry that must be balanced against privacy requirements and bystander consent.

1.4 Shifts to private/semi-private spaces

“ For adversaries, private and semi-private spaces enable high-trust manipulation and lower-risk iteration.

More interactions are moving into private and semi-private channels, such as encrypted messaging, invite-only groups, broadcast channels and subscriber communities, which is shifting where coordination, persuasion and harm formation take place. WhatsApp alone has reported more than 3 billion monthly users, illustrating the scale of private messaging as a primary communication layer.[21] These environments can feel more trusted and context-rich, and some observability constraints within them, such as end-to-end encryption and privacy-by-design choices, are legitimate security and privacy features rather than defects. At the same time, they reduce provider visibility, requiring teams to operate within these limits using structural signals, metadata and external reporting.[22] As a result, harmful activity is increasingly distributed, it may be organized in closed spaces, tested in smaller communities and then operationalized elsewhere, with spillover effects across multiple services, intermediaries and organizations that may never host the original coordination.[23]

For adversaries, private and semi-private spaces enable high-trust manipulation and lower-risk iteration. Within these spaces, scams spread through forwarded messages, while recruitment and coercion take place away from public visibility; and narratives can be tailored to specific communities with rapid feedback.[24] Evidence often circulates as screenshots, short clips or paraphrased posts that are difficult to authenticate and easy to weaponize, while moderators and responders have fewer signals to use and a higher degree of uncertainty. This dynamic also supports hand-off tactics where an initial lure or grooming interaction occurs in a closed channel, following which the victim is moved to a different surface for payment, credential capture or exploitation, complicating attribution, enforcement and victim support across organizational boundaries.

The key shift is that prevention and response rely less on content visibility and more on structural signals, user and admin controls and cross-service coordination, which requires measurement and decision frameworks that differ from public-feed moderation while remaining effective under constrained observability.

1.5 Regulatory evolution

In Europe, the Digital Services Act (DSA) moves platforms towards continuous, risk-based governance. Large services must map systemic risks, implement mitigations, undergo independent audits, enable researcher access, increase recommender transparency and user controls, verify marketplace traders and support crisis response mechanisms.[25] The EU AI Act combines a risk-tiered regime containing prohibitions for some uses, strict duties for high-risk applications and a dedicated approach for general-purpose AI with transparency for synthetic media and AI interactions, driving durable logging, evaluation pipelines and auditable safety practices.[26] The UK Online Safety Act (OSA) requires in-scope user-to-user and search services to assess illegal-content risks and adopt proportionate safety-by-design measures under the codes issued by the United Kingdom Office of Communications (Ofcom), prioritize child protection with robust age assurance and meet strengthened transparency and appeals requirements – backed by significant penalties and business-disruption powers.[27]

Beyond Europe, Australia has had an Online Safety Act in place since 2015, recently implementing a novel social media minimum age obligation requiring "reasonable steps" to prevent under-16s from holding accounts on covered platforms.[28] Singapore's Infocomm Media Development Authority (IMDA) enforces the Code of Practice for Online Safety on designated platforms requiring risk controls and annual reporting.[29] There is also South Korea's AI Basic Act, Brazil's Children and Adolescents Online Safety Act (Estatuto da Criança e do Adolescente – Digital ECA) and the United Arab Emirates' Child Digital Safety Law and Online Safety Act.

Fragmentation creates operational complexity for organizations and an opportunity space for adversaries. Definitional divergence affects foresight directly. When jurisdictions define the same harm differently, whether it is varying age thresholds, different scoping of AI-generated content or divergent standards for what constitutes illegal or harmful material, the categories for which an organization scans may reflect one jurisdiction's framing and miss patterns already recognized elsewhere. At the same time, stronger transparency, auditing and risk-management requirements can improve the shared evidence base if implemented in ways that support learning and mitigation effectiveness rather than performative reporting and are realistically achievable. The Global Online Safety Regulators Network has noted the compliance challenge and looks to pursue regulatory coherence to ensure safety does not stop "at the border" and that companies can benefit from legal certainty.[30]

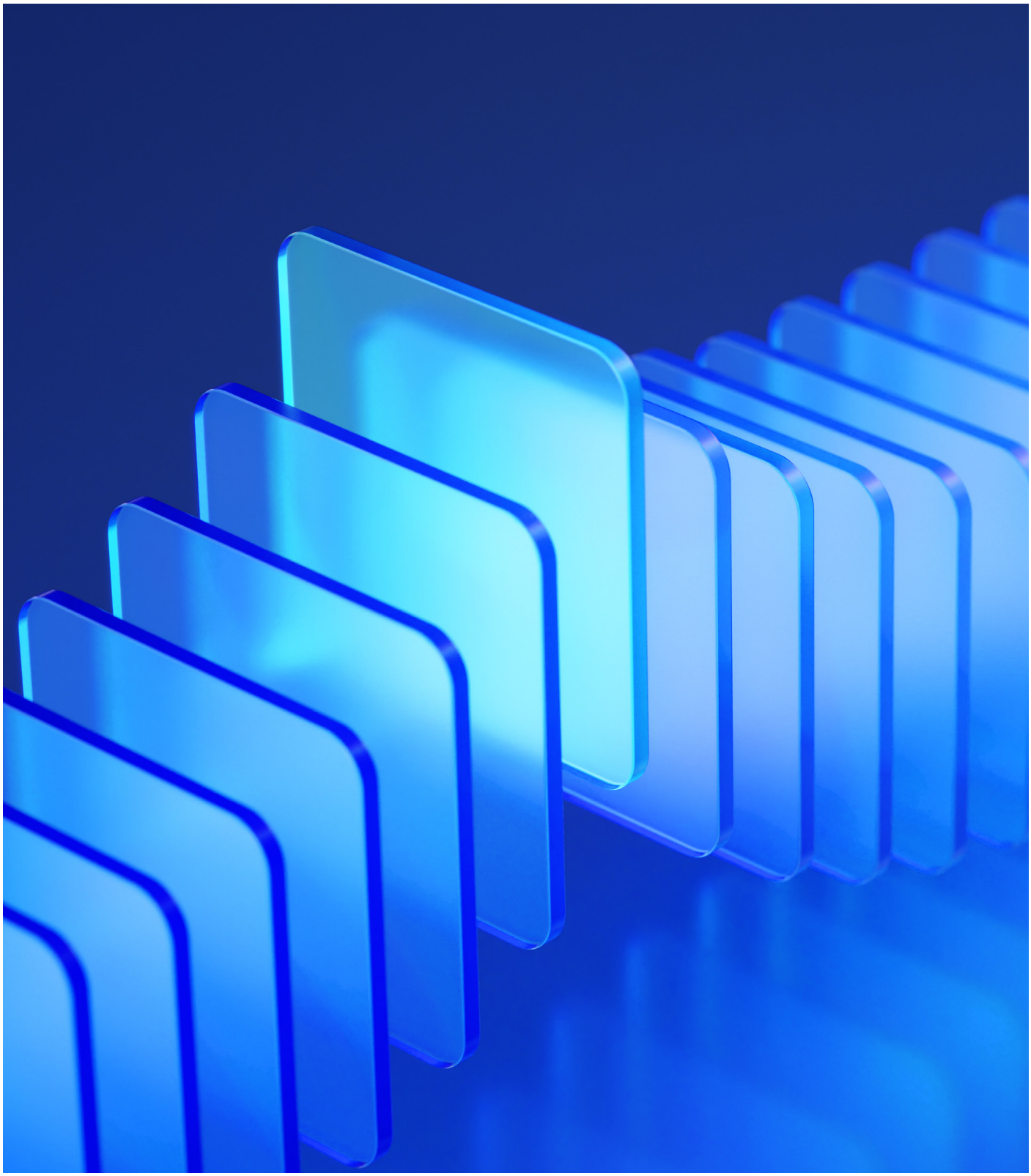
Regulatory fragmentation is not limited to online safety and AI laws. Organizations increasingly operate across a patchwork of rules affecting data governance, identity, advertising and privacy, each with different definitions, timelines and enforcement models. Compliance obligations may lead companies to change products and processes to age assurance, identity verification, expanded transparency reporting and jurisdiction-specific feature constraints that can shift where harms occur and the evidence is available to investigate them. Comparative reviews of online safety regulation show significant divergence across jurisdictions in content-based, design-based, transparency and procedural requirements, raising the cost and complexity of operating consistently across markets.[31]

Adversaries exploit these seams. Limits on cross-border data sharing can weaken both early warning and enforcement by preventing signal aggregation, slowing investigation and enabling actors to launder activity through jurisdictions with weaker cooperation.[32] Compliance mechanisms themselves can also become attack surfaces, with age-gating and verification flows creating new opportunities for phishing, identity theft and synthetic-identity circumvention, while differing age thresholds and checking requirements can push users towards less trustworthy services that ignore safeguards.[33] Regulations should be treated as a safety variable, not only as a legal one.

❝ Differing age thresholds and checking requirements can push users towards less trustworthy services that ignore safeguards.

2 Digital safety foresight pillars

By deploying foresight in digital safety operations, organizations can be ready to act quickly and pre-emptively against adversaries.



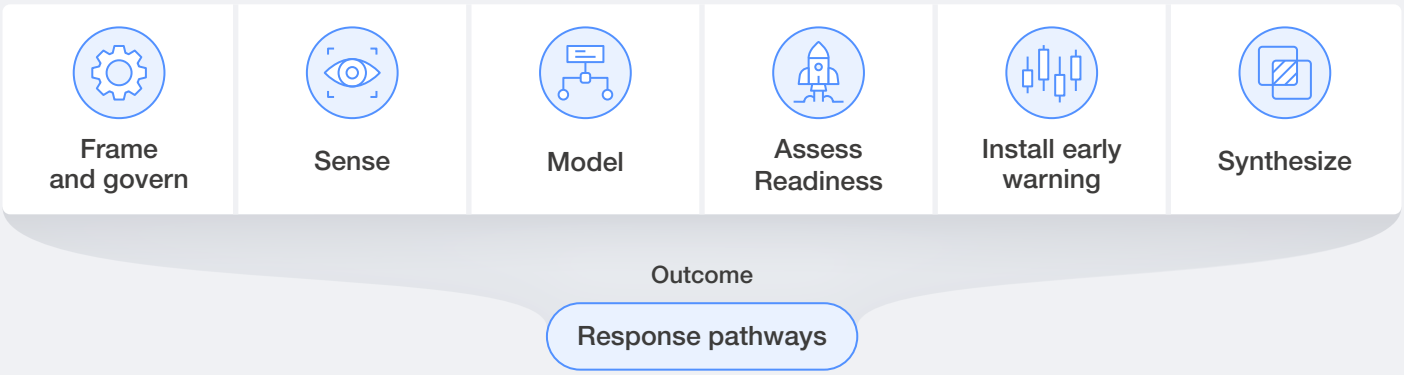
This section presents a set of foresight pillars for digital safety operations based on best practices that can help organizations stay ahead of emerging harms rather than reacting only after an impact is widespread. The pillars translate signals and early concerns into structured, decision-ready outputs.

Because most organizations already assess risks arising from their own product changes due to internal policies or external regulations, these pillars are intentionally tuned to focus on harms driven by exogenous change: shifts in technology, politics, markets, culture or adversary behaviour that occur outside the product roadmap but that leverage existing product capabilities in new or harmful ways. Endogenous risks are not ignored; they are incorporated where they interact with these external drivers or create new exposure.

The pillars culminate in a concise harm card that feeds directly into section 3. Placement in the response pathways in section 3 is based on information from the foresight pillars.

To make the pillars concrete, the same illustrative example is used throughout: commodity AI impersonation kits go mainstream, with actors using AI voice/video and automated agents to impersonate trusted contacts in real time across channels, enabling scalable fraud, coercion/harassment and access abuse.

FIGURE 1 **The six digital safety foresight pillars**



Source: World Economic Forum

2.1 Frame and govern

Frame and govern is the pillar that makes digital safety foresight decision-relevant and accountable. It sets the question, boundaries and decision rights so that later analysis translates into action.

Decision brief and scoping question

Many organizations anchor the work in a short decision brief that acts as a scoping question. The brief clarifies the objectives, defines the time horizon and frames the work so it does not sprawl into general trend-watching. Once agreed, the scoping question should frame every method used within the pillar set, and decision-makers should be involved early and throughout so the output becomes a change in posture. Framing should also be explicit about perspectives, such as whose lived experience, regional context and domain expertise needs to be represented to surface blind spots and internal assumptions.

A strong brief is concise but explicit. It typically records:

- The decision(s) the organization is preparing to make and who owns them.
- The planning horizon(s) that matter (for example weeks for crisis preparedness and months for product exposure).
- What geographies, languages, modalities and user or customer contexts are in scope.
- What would count as meaningful change in risk posture.
- Any jurisdiction-specific definitions that materially alter the harm taxonomy or response obligations.

> Governance is usually cross-functional: the goal is not micromanagement but accountable alignment on risk appetite and response posture.

The brief should also distinguish exogenous drivers outside organizational control from endogenous drivers created by internal choices, because the response options, incentives and accountability differ.

Operating context and assumptions

Digital safety foresight often fails when it assumes a particular service shape, such as direct or indirect user relationships, full content visibility or a single enforcement surface. Framing should therefore capture the operating context upfront, including what is observable, where responsibilities are shared across partners, customers and third parties, and what levers exist, such as product design constraints, suspicious sign-ups or abuse, customer notifications, account or asset suspension, takedown processes, contractual controls, friction and throttles, policy and enforcement, etc. These constraints change what evidence is available, which early signals are realistic and which actions are credible. Foresight should be designed to work within the constraints.

Assumptions should be treated as working hypotheses, not hidden premises. A light uncertainty or assumptions register helps teams track what they are relying on, what would disconfirm those assumptions and which assumptions would materially change the decision if wrong.

Cadence, ownership and documentation

Framing is not complete without ownership. If no one is accountable for acting on the output, foresight degrades into analysis. Governance should assign clear decision rights (who can accept risk, who can change product or policy, who can trigger incident-style response) and clarify the organization's risk posture for digital safety, such as what trade-offs are acceptable and what escalation thresholds apply when evidence is thin but stakes are high. Because digital safety spans trust and safety, security, privacy, legal and product strategy, governance is usually cross-functional: the goal is not micromanagement but accountable alignment on risk appetite and response posture.

Governance is sustained through repeatability with minimal overheads. A practical scaffold includes: a scan log that captures what was noticed and why; an assumptions register that records uncertainty and what would change the view; and a decision log that records what was asked, decided, deferred and why. These documents should be lightweight and revisable so teams can update them. This creates organizational memory that is legible and auditable.

BOX 1

Example harm: Frame and govern pillar

The work would begin by framing "commodity AI impersonation kits" as a concrete, decision-facing risk. This includes whether and how to change posture around the specific moments synthetic impersonation most reliably exploits – such as trust cues and identity assertions, account recovery, admin or permission changes – and any high-stakes actions, such as transfers, payouts or sensitive data access.

The decision brief would set the boundaries that matter for this harm, such as which surfaces are in scope, which geographies/languages and user/customer contexts are most exposed and which outcomes are the priority (financial fraud, coercion, account or privilege compromise, etc.). Assumptions would also be recorded that will drive later work – for this example, attacker access to voice/video samples, likely targeting of admins/high-trust roles, expected channels for first contact and anticipated scaling tactics, such as multichannel laundering.

Because evidence will differ by service shape, the framing would explicitly note what can and cannot be observed and name the levers available, then establish a clear escalation rule for high-stakes uncertainty – for example, external reporting that playbooks are targeting the organization's brand or a cluster of impersonation-linked recovery attempts, so the organization can move into deeper modelling, readiness checks and early warning without waiting for fully confirmed impact.

2.2 Sense

> Scanning is most robust when treated as a collective intelligence that actively convenes and incorporates multiple perspectives.

Sense is the pillar that expands an organization's field of view beyond what is already debated internally or presently visible through day-to-day operations. It includes a disciplined form of horizon scanning – a structured, evidence-gathering practice that identifies early signals of change that could plausibly create new vectors for harm.

Sense is not about forecasting a single future. It is about noticing change early enough to:

- Challenge assumptions
- Clarify what could become materially important
- Feed decision-relevant analysis under the other pillars

Sense scans widely to detect emerging conditions, new capabilities and shifting incentives.

What sense is looking for

Digital safety organizations operate across many types of platform, system and product model with direct and indirect relationships, high and low visibility, content and non-content surfaces, consumer and enterprise contexts. Sense therefore focuses on types of change that travel across contexts.

In practice, scanning often looks for shifts in:

- **Capabilities:** Tools and techniques that reduce the cost of abuse (automation, synthetic media, translation, domain provisioning, scaling infrastructure, identity spoofing).
- **Incentives:** Monetization routes, market demand, enforcement dynamics and regulatory or business shifts that change the attractiveness of harm.
- **Access and distribution:** Integrations, ecosystems, tenant creation, defaults, intermediaries and new channels that change reach or scalability.
- **Evasion and adaptation:** New laundering patterns, testing behaviours, account/IP cycling and "workarounds" that indicate that adversaries are iterating.

Signals

Sense should include sources that reflect the service's domain, adjacent domains where harms migrate and macro shifts that reshape the environment. Rather than a universal list, many organizations structure sensing inputs as external and internal signals.

External signals might include research and civil society reporting, investigative journalism, practitioners, regulatory and standards developments and cross-industry patterns. To make these signals usable, organizations should establish predictable external intake channels – for example, a dedicated email or web form for trusted reporting, a standing point of contact for researchers and civil society and clear guidance on what information is most helpful. For time-sensitive issues, a fast path should be available for trusted partners to flag urgency and reach a rapid triage process backed by internal protocols. External actors should be updated on their submissions when possible.

Scanning is most robust when treated as a collective intelligence that actively convenes and incorporates multiple perspectives through recurring check-ins, partner roundtables or researcher engagement.

Internal signals can include trust and safety operations data (clustered reports, appeals, edge cases), security/privacy indicators (abuse-of-service precursors, access anomalies, recovery spikes), customer/admin channels (tickets or escalations) and product/ecosystem observations (unexpected usage patterns and stress points). The specific signals differ by context, but what matters is that the pillar recognizes where signals are strong or indirect and where proxy signals must stand in (due to lack of direct signals).

It is also important to define baselines and simple triggers. For each indicator, agree on what normal looks like, then define simple trigger conditions, such as how big a deviation, how fast or how long an elevation would be concerning. The focus is less on perfect statistical thresholds and more on shared, documented rules of thumb.

Confidence in measurements is critical, as it means knowing how much to trust the picture being drawn. However, understanding and measuring digital safety outcomes remains complex. A previous report from the Coalition, *Making a Difference: How to Measure Digital Safety Effectively*, 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.

Although the actual metrics will vary, these categories can help sort them.

Interpreting signals

Sense adds value when it produces interpretable outputs that are short, coherent insights that explain what is changing and why it could matter rather than accumulating a long list of references.

An insight record is typically anchored in:

- **The change:** What is newly observable or newly plausible.
- **Why it matters:** The potential implications for digital safety harms, including abuse-of-service and privacy-linked harms.
- **Conditions and assumptions:** What must be true for the change to become material.
- **Evidence quality:** What is known, what is inferred and what is uncertain.
- **Time horizon:** Whether this is a near-term operational pressure or a longer-term structural driver.

Sense should be explicit about the operating constraints that shape evidence quality; these constraints (especially those related to privacy) should not automatically be treated as deficits to be eliminated but rather seen as checks to be incorporated/worked with. Weak signals are inherently ambiguous; additionally, some services operate with limited visibility. Such constraints do not prevent sensing, but they do increase the value of triangulation across sources. Organizations should avoid treating hard-to-observe signals as being of low importance or mistaking partial visibility for full certainty and attach a simple confidence note that separates what is observed from what is inferred.

BOX 2

Example harm: Sense pillar

Early signals of commodity AI impersonation could include rapid diffusion of low-cost voice/video-cloning tools, mainstream reporting of impersonation-enabled fraud or harassment and early policy attention around biometric likeness, consent and authentication. To act within the Sense pillar, an organization could deliberately pull these external signals into its scan log through a curated watchlist and via an external intake route, so that credible concerns are not missed. Internally, the team would look for converging themes and appropriate proxy indicators, such as clusters of impersonation-related reports, spikes in account recovery attempts, unusual patterns around high-trust or high-stakes workflows, support or customer/admin escalations about verification or complaints suggesting non-consensual likeness.

2.3 Model

Model is the pillar that turns signals of change into decision-relevant explanations of how harms could emerge, scale and adapt. Where Sense broadens the field of view, Model narrows it into structured thinking about mechanisms of the plausible pathways through which capability, incentives, access or evasion could translate into harm.

Model helps teams build a shared understanding of direct or indirect relationships, high or low observability, privacy-protecting modalities and multiparty ecosystems by making assumptions and uncertainties explicit.

Key aspects to capture in the Model pillar:

- **Actors and motivations:** Who might drive the harm, including opportunistic abuse and organized actors, what they want and what constraints they face.
- **Enabling conditions:** What must be true for the harm to scale, such as feature affordances, defaults, ecosystem access, social conditions, business incentives.
- **Paths and leverage points:** How the harm unfolds across stages and where it can be interrupted from design friction, policy boundaries, enforcement levers and admin controls to ecosystem interventions.
- **Adaptation and evasion:** How a motivated actor could route around controls, shift targets or migrate to new channels as conditions change.
- **Impacts and distribution:** Who is affected and how, including differential impacts on at-risk groups and whether harms are transient, cumulative or irreversible.

Actors and motivations

Understanding who may drive a harm, and why, is foundational, because it determines a harm's speed of scale, how persistent it is and what forms of deterrence or friction can affect it. Actor analysis also helps avoid modelling harms as though they arise from general misuse, when in practice they are often driven by a small number of motivated actors who iterate, coordinate and reuse methods across systems.

To capture actors and motivations, organizations often use a simple actor map that describes a few plausible actor categories and what constrains them. This is not about exhaustive threat intelligence; it is about recognizing operating models – such as opportunistic, organized, insider, ideological, financially motivated or coercive – and how those models interact with the organization's relationship structure with direct users, admins, customers, partners and ecosystems.

Ways to capture this:

- **Actor–goal–capability sketch:** For each harm mechanism, capture:
  - Likely actor types
  - Their goals
  - Their capabilities and resources
  - Their constraints/costs
  - What success looks like

Enabling conditions

Enabling conditions describe what must be true in the product, ecosystem or external environment for a harm pathway to become feasible and scalable. This lens is where teams separate the possible from the material, because harms rarely emerge in a vacuum; they emerge when capabilities, incentives, access and gaps align.

Such conditions should include both system affordances – including design choices, defaults and workflows – and context affordances, such as market dynamics, intermediaries, geopolitical shifts and social norms. It is additionally important to account for an organization's real constraints, from observability to reliance on partners.

Ways to capture this:

- **Condition checklist:** Identify between three and seven conditions that would make the harm materially more likely, such as low-friction onboarding, weak trust cues, high-privilege workflows and low-cost scaling.
- **Dependency mapping:** Note which external systems or intermediaries have the biggest impact on the harm, such as integrations with other organizations, marketplaces and payment and identity providers.

“Teams separate the possible from the material, because harms rarely emerge in a vacuum; they emerge when capabilities, incentives, access and gaps align.

Paths and leverage points

Pathway modelling captures the sequence of events and decisions through which a harm unfolds, including where it can be interrupted. It is often the most operationally useful output because it translates abstract risk into concrete points. Critical leverage points include product choices, friction points, policy boundaries, enforcement steps, partner coordination and response triggers.

Ways to capture this:

- **Causal pathway map:** Define a start state, intermediate steps and the resulting harms; annotate each step with "what enables this step" and "what could block it".
- **Bow-tie:** Choose the pivotal point at which control is lost, map precursors on the left and consequences on the right then list the barriers on both sides.
- **Leverage prompts:** Identify two to four moments at which a small intervention would likely change the outcomes disproportionately – for example, defaults, irreversible actions and high-trust cues.

Adaptation and evasion

Adaptation modelling anticipates how actors may respond once mitigations are introduced, such as how they route around controls, shift to adjacent surfaces, change target selection or exploit secondary effects. Without this lens, readiness work may be effective for the first wave of abuse but fragile against iteration. Evasion does not require the publishing of operational how-to details; rather it requires identifying plausible bypass categories (or example, account cycling, multichannel laundering, testing and probing, trust-building "warming", abuse of legitimate workflows) and the likely second-order effects of mitigations, including displacement to new channels and increased targeting of vulnerable groups.

Ways to capture this:

- **Evasion tree:** For each key barrier, list three to five plausible bypasses (substitution, timing, migration, exploitation of edge cases).
- **Tactic-style mapping:** Describe how actors gain access, establish trust, scale and evade; keep it at the behavioural level.

Impacts and distribution

Impact modelling clarifies who is harmed, how severely and how harms proliferate. This lens matters because the same mechanism can produce different outcomes depending on who is exposed, who is vulnerable and who has capacity to cope or recover.

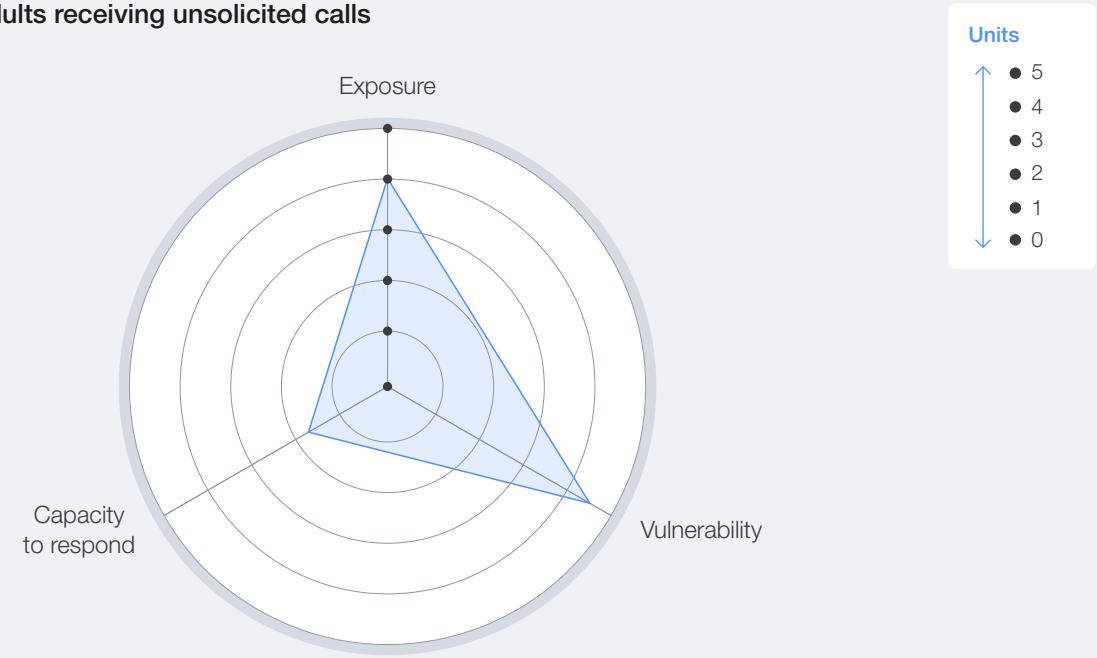
A useful framing is to separate the likelihood that a harm pathway will be exercised from its impact. To represent distribution, organizations can adapt the vulnerability–exposure–capacity (VEC) style of thinking common in risk and resilience work. Vulnerability reflects susceptibility to harm; exposure reflects who is likely to encounter the hazard; and capacity reflects ability to anticipate, cope and recover.

Ways to capture this:

- **VEC radar:** For priority groups or contexts, plot vulnerability, exposure and capacity on a simple radar chart to visualize uneven risk distribution (see Figure 2).
- **Impact ladder:** Classify impacts by severity, reversibility and accumulation.
- **Distribution prompts:** Who is disproportionately targeted; which users/customers have fewer protective resources and where reporting and redress are weaker.

FIGURE 2 | **VEC radar for conversation impersonation fraud**

Example group: Older adults receiving unsolicited calls



Source: World Economic Forum

BOX 3 | **Example harm: Model pillar**

An organization would translate "synthetic impersonation is diffusing" into a mechanism capture that explains how it would plausibly work in this service and where it can be disrupted.

This includes:

- Actors and motivations (for example, opportunistic fraudsters, organized scam rings, targeted harassers/coercers) and what success looks like.
- Enabling conditions, such as easy access to voice/video samples, weak trust cues and high-stakes workflows susceptible to social engineering, including recovery, admin/permission changes, payouts/transfers and sensitive data access plus relevant dependencies.
- A pathway and leverage points map (for example, target selection → identity capture → synthetic persona creation → trust-building contact → trigger action [transfer, credential reset, admin change] → scale via automation). Highlighting two to four leverage points where friction or verification could break the path.
- Adaptation/evasion expectations captured as a short barrier-bypass note (for example, multichannel laundering, account warming and threshold testing).
- Impacts and distribution, making a quick VEC analysis to flag who is most exposed and least able to verify or recover.

2.4 Assess readiness

Assess readiness is the pillar that tests whether an organization can credibly prevent, detect, respond to and learn from digital safety harms. It asks whether if the harm pathways described in the Model pillar began to materialize, would there be enough coverage and capacity to reduce impact and to do so consistently for the organization's service not to be affected?

The purpose is not to grade an organization against an ideal standard, but to clarify where coverage is strong or thin. This is important because practices that work well for one type of organization may be ineffective or create unintended consequences in another, with inflexible approaches rarely translating well.

> Assess readiness is not only a set of controls but a capability to coordinate decisions and execution across teams and external dependencies.

Surfaces mapped

Readiness should be evaluated relative to the organization's actual levers and constraints. Rather than assuming one enforcement surface, Assess readiness looks across the control surfaces, which may include:

- **Product and design levers:** How the system is built, defaults, friction and user/admin journeys.
- **Policy and governance levers:** Rules, standards and decision rights that determine what the organization is willing to tolerate and who can act.
- **Operational and response levers:** Escalation pathways, response playbooks, communications, personnel expertise, redress and coordination mechanisms that translate a risk posture into action.
- **Ecosystem and external levers:** Partner policies, contractual terms, marketplace mechanisms and coordination with externals where internal levers are limited.

Coverage

Assess readiness is strongest when it can demonstrate coverage across the harm's pathway.

A readiness assessment therefore examines whether the organization has credible capability in three areas:

- **Prevention:** Are there design choices, defaults or safeguards that reduce the likelihood of harm or slow its scaling?
- **Detection:** Can the organization notice the harm early enough through direct signals or proxies, and can it interpret what it is seeing amid uncertainty?
- **Correction:** Is there a repeatable way to capture what happened, update assumptions and improve controls without waiting for a crisis? Can the organization act quickly, consistently and with clear accountability, including cross-functional alignment?

Assess readiness is not only a set of controls but a capability to coordinate decisions and execution across teams and external dependencies, especially where the organization has no direct control and must design for adaptation. It should test whether the organization can act when evidence is unclear but potential impact is high.

Using a coverage and control map can help identify the gaps in a consistent way. Coverage should be treated as conditional where a control depends on customer enablement and response or third-party action. A control the organization can offer but not activate or enforce is not equivalent to one it owns.

TABLE 1 **Template of a coverage and control map**

Harm pathway stage	Controls in place (policy/ product/ model/operational/ human/user/ external)	Control type (P/D/C)*	Ownership/ dependency (fully owned, shared, customer-enabled/ customer-dependent, third party-dependent)	Where it applies (surface/ region/ language/ group)	Evidence it works (incidents, audits, evaluations)	Gaps/ failure modes
Enabling conditions						
Attempted misuse						
Harmful event						
Spread or escalation						
User response and recovery						

* P = preventive, D = detective, C = corrective

Evidence

Assess readiness is strongest when the organization has some way to validate whether its barriers and processes behave as intended, particularly once actors adapt. Assurance does not have to be heavy. Many organizations use lightweight practices, such as:

- **Scenario walk-throughs/tabletop exercises:** To reveal decision bottlenecks, ambiguity and coordination gaps.
- **Control "break tests":** Targeted review of known weak points (high-leverage workflows, ambiguous policy edges, partner escalation paths).
- **Unintended consequence checks:** Whether safeguards displace harms to harder-to-see surfaces, increase the burden on users/customers or disproportionately affect certain groups.

This part helps prevent the adoption of the assumption that having some controls implies readiness, when the real constraints are speed, authority, evidence quality and cross-team coordination.

BOX 4 **Example harm: Assess readiness pillar**

Using the mechanism capture from the Model pillar, the organization would test whether it has credible coverage at the leverage points that impersonation exploits – and whether that coverage holds under constraints.

This means mapping the impersonation pathway across stages of enabling conditions, attempted misuse, harmful event, spread/escalation and user/customer recovery. This step should be followed by completing a coverage and control map.

For impersonation, the readiness review would concentrate on high-impact moments, such as trust cues and identity assertions, admin/permission changes, payouts or other sensitive actions. This would then be stress tested to determine whether detection is viable using the signals available.

Finally, it would evaluate execution readiness. The output is a clear statement of readiness, containing constraint notes, the most consequential gaps and failure modes and a small set of owned improvements that will materially reduce impact.

2.5 Install early warning

Install early warning is the pillar that translates modelling into ongoing awareness and support efforts in ensuring readiness. Its question is therefore not "what is changing in general?" but "are conditions for this harm shifting enough for us to change posture?" The purpose is to reduce strategic surprise and decision latency by monitoring a small set of meaningful watchables linked to the harm mechanism, the relevant control points and clear escalation routes. It is not about perfect prediction or comprehensive measurement. It is about seeing enough sufficiently early to act proportionately.

Monitoring

Many organizations build an indicator portfolio that covers four categories:

- **Environmental signposts:** External developments that alter the harm landscape, such as tool diffusion, market formation, regulatory shifts, migration patterns and major geopolitical events that change targeting.
- **Actor and capability signposts:** Observable evidence that relevant actors are experimenting, coordinating or scaling changes in tactics and indicators of professionalization.
- **System and control signposts:** Stress signals that the organization's controls are being probed or strained due to unusual patterns around high-leverage workflows, sudden spikes in edge cases, suspicious provisioning, increased bypass attempts or rising false positives/appeals.
- **Harm confirmation signals:** Reports, complaints, tickets, user/customer escalations and other corroboration that impact is occurring.

Because many emerging harms are initially difficult to observe directly, monitoring should combine quantitative and qualitative inputs. Internal telemetry, support data, moderation queues, fraud operations, red-team findings, trust and safety reviews, partner reports and external expert reporting should be considered where relevant. Often, the earliest warning will come not from a clean metric but from an unusual pattern noticed by front-line teams or external partners through behavioural patterns, usage shifts or novel forms of manipulation, as they have direct exposure. Monitoring should include a clear way to incorporate non-standard observations rather than treating only structured data as legitimate evidence.

> The purpose is to reduce strategic surprise and decision latency by monitoring a small set of meaningful watchables linked to the harm mechanism.

Indicators

Indicators should be chosen because they are likely to move if the harm is activating or if a control is failing, and because the organization can observe them at the right cadence to intervene.

A practical method is to create watchables:

- **Actors and motivations:** Signals of motivation and scale (monetization routes, volume and coordination proxies, targeting patterns).
- **Enabling conditions:** Signposts that those conditions are strengthening (reduced capability cost, new distribution channels).
- **Pathways and leverage points:** Indicators at high-leverage moments (identity assertions, high-stakes actions, high-reach distribution).
- **Adaptation and evasion:** Indicators of bypass attempts or migration (new workarounds, channel switching, laundering).
- **Impacts and distribution:** Disproportionality signals (who is affected, severity, reversibility, concentration in segments).

Indicators should include both thresholds and patterns. Some harms will show through clear spikes; others will show through clustering, persistence over time or multiple signals moving together. Where baselines are new and still developing, organizations should use interim reference points, such as rolling averages or related harm categories, while noting the limitation. A weak baseline is often better than no watchable.

Teams often operationalize this by asking two questions:

1. What would likely be seen first if this pathway starts? (the earliest observable precursor)
2. What would need to be seen to justify changing posture? (the confirmation threshold)

This keeps Install early warning tied to decision relevance rather than to whatever is easiest to count.

Indicators should also be disaggregated where possible. A mild aggregate shift may hide a serious concentration in a particular geography, workflow or exposed group. Indicators should be refined over time as teams learn.

Tripwires and sentinels

For some harms, organizations will not have strong-enough indicators to wait for conventional evidence. This is especially true where harms are novel, difficult to measure, highly consequential or likely to scale quickly. In these cases, organizations should use tripwires and sentinels.

Tripwires are pre-agreed conditions for precautionary action. A tripwire should therefore not be framed as proof that the harm is occurring at scale, but rather as a decision rule for when uncertainty has become sufficiently concerning for a stronger posture to be necessary. A tripwire should specify the condition or combination of conditions that matter, who has authority to activate, what action should follow, whether that action is reversible, how quickly the decision will be revisited and what additional evidence would support escalation or stand-down. The strongest tripwires are usually not based on a single datapoint but on combinations.

Tripwires are often difficult to calibrate. If they are too insensitive, they will rarely trigger until there is already ample evidence; if they are too sensitive, they can create unnecessary disruption and erode confidence. The practical response is to keep them limited in number, tied to consequential harms and paired with time-bounded, reviewable actions. In other words, tripwires should not trigger open-ended panic measures but defined precautionary steps that can be re-evaluated quickly.

Sentinels provide the human detection layer. They are the people and structures that notice, interpret and elevate signals that do not yet fit established metrics or categories. Sentinels need a mandate and a route to escalation that will lead to review. This means identifying who the relevant sentinels are, giving them a method for escalation, creating a rapid triage forum that can assess signals without excessive procedure and ensuring that concerns are recorded to be compared later to observations.

BOX 5 | **Example harm: Early-warning pillar**

Once the impersonation mechanism is modelled and the leverage points identified, the organization should create an early-warning card consisting of a short set of signposts to monitor, with ownership, and what changes are triggered at different thresholds. The monitoring set would be deliberately narrower than Sense and tied to the first observable precursors and high-impact signals.

For example, the team might track:

- External diffusion signposts (growth in low-cost cloning, new widely shared playbooks, spikes in public reporting).
- Internal precursor signals at leverage points (unusual recovery attempts, new-device plus sensitive action sequences, repeated probing of verification/trust cues).
- Control strain signals (surges in impersonation-related tickets, appeals or moderation backlog).
- Impact confirmations (clusters of reports naming impersonation, verified fraud/coercion cases, disproportionate targeting of high-exposure groups).

The organization would then follow predefined actions aligned to the harm and its severity, until a more sustained, fully developed response is in place at the response pathways.

2.6 Synthesize

Synthesize is the pillar that consolidates outputs from the other pillars into a single, decision-ready view. Its purpose is to present a coherent picture. Synthesis brings together the scope and constraints captured in Frame and govern; the interpreted signals and insight records from Sense; the harm mechanism from Model; the coverage and gaps identified in Readiness; and the signposts and triggers defined in Install early warning. The result should be a compact document that leadership and operational teams can use to select a response pathway.

A common synthesis output is a harm card (or equivalent briefing note) that is short enough to be read quickly but structured enough to preserve reasoning and traceability (see Figure 3 for an example).

A harm card can include:

- A plain-language description of the emerging issue and why it matters.
- A mechanism summary including the actor types, enabling conditions, pathway/leverage points and likely adaptation.
- Who is most affected and where impacts concentrate.
- Operating constraints.
- A readiness snapshot of whether coverage is thin/partial/strong plus the most consequential gaps/failure modes.
- The early-warning indicators.

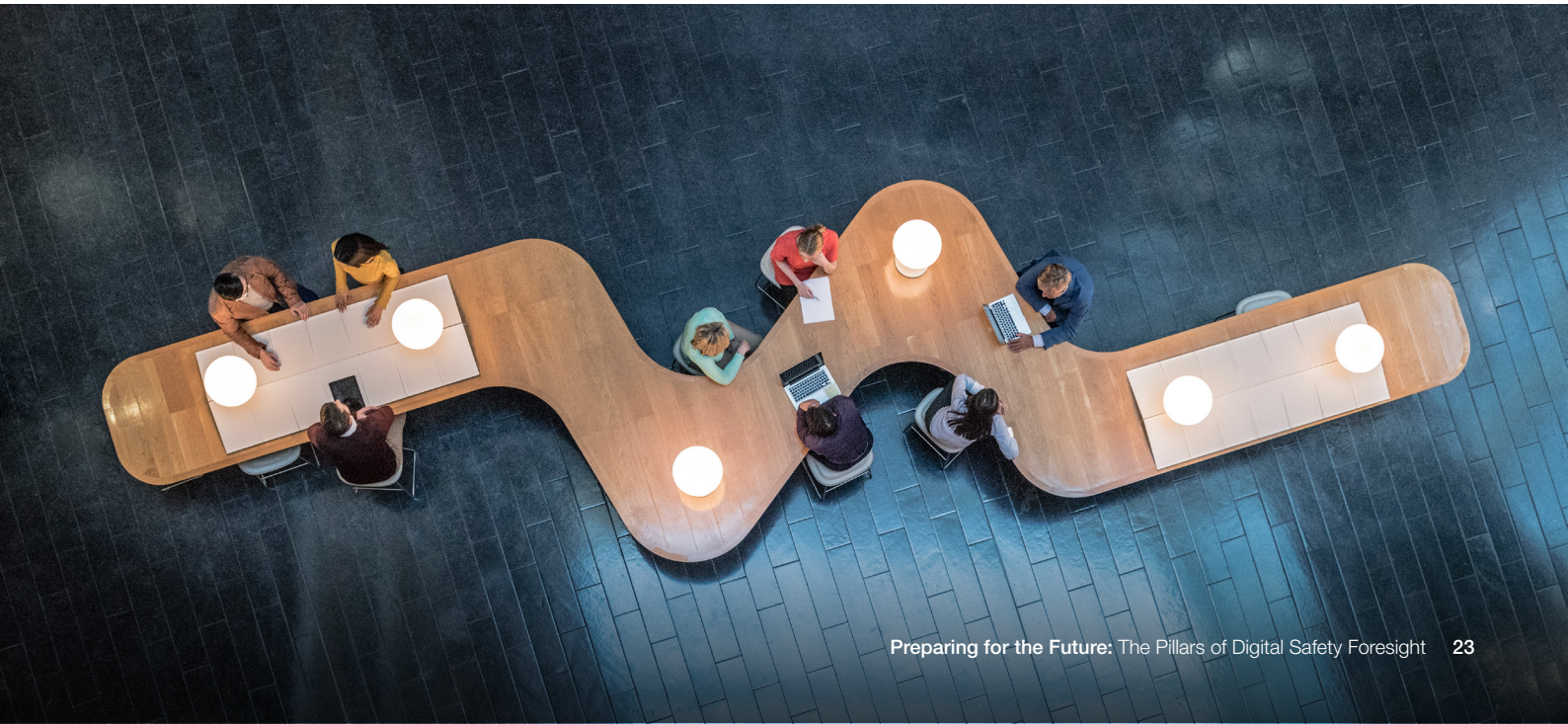


FIGURE 3 | **Harm card example: Commodity AI Impersonation**

Harm card-01: Commodity AI impersonation at scale

Tier: 1 – Critical

Actors use AI voice/video and agents to impersonate trusted contacts in real time, enabling scalable fraud, coercion and access abuse.

Plain-language issue and why it matters

Lowers the cost of targeted social engineering and increases scale across channels

Exploits high-stakes workflows (recovery, admin changes, payouts, sensitive access)

Who is most affected and where impacts concentrate

High-trust roles (admins, sellers, community leaders) targeted for access and high-stakes actions

Older adults/low-digital-literacy contexts facing unsolicited calls/messages

Diaspora and remittance contexts – low-trust settings with limited verification capacity

Mechanism summary

Actors:
Opportunistic fraudsters; organized scam rings; targeted harassers/coercers

Enabling conditions:
Public samples; weak trust cues; low-friction, high-stakes workflows

Pathway and leverage points:
Target → sample → clone → trust contact → trigger action; disrupt at recovery; admin change; payouts; trust cues

Likely adaptation:
Multichannel laundering; account warming; threshold testing

Operating constraints (what shapes evidence and action)

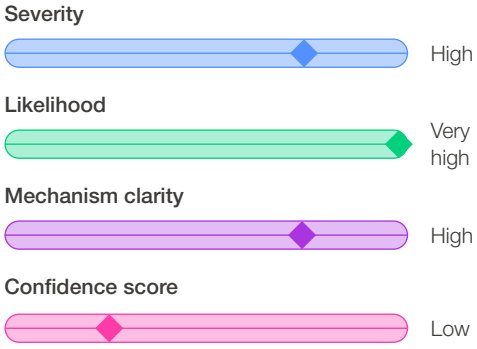
Visibility:
Content/metadata/aggregates (often partial; proxies needed)

Relationship model:
Direct users and/or tenant/admin-mediated workflows

Control surfaces:
Trust cue; friction and throttles; enforcement; admin controls

External dependencies:
Telcos/messaging routes; identity providers; payment rails; partners

Ratings (non-prescriptive)



Assess readiness snapshot

Coverage:
Partial

Key gaps: Recovery/admin-change friction and step-up verification; detection leans on late signals and limited cross-channel visibility

Install early warning and assumptions

Early warning indicators:

- → External diffusion of kits/playbooks; sector incidents
- → Spikes in recovery attempts; admin/permission changes
- → New-device + sensitive action sequences; clustered reports/tickets

Assumptions and evidence:

Assumptions:
Samples accessible; workflows engineerable at scale

Stronger:
Confirmed incidents/probing/partner reports

Weaker:
No precursor-impact linkage; mitigations reduce attempts without displacement

Routing to action pathways

Proposed pathway selection:
Pathway 2 – Novel/uncertain coverage (mechanism plausible; coverage partial; confidence limited by visibility and adaptation)

Linked hypotheses: Fraud; coercion/harassment; privilege compromise

Source: World Economic Forum

FIGURE 3 | **Harm card example: Commodity AI Impersonation**

Harm card-01: Commodity AI impersonation at scale

Tier: 1 – Critical

Actors use AI voice/video and agents to impersonate trusted contacts in real time, enabling scalable fraud, coercion and access abuse.

Plain-language issue and why it matters

Lowers the cost of targeted social engineering and increases scale across channels

Exploits high-stakes workflows (recovery, admin changes, payouts, sensitive access)

Who is most affected and where impacts concentrate

High-trust roles (admins, sellers, community leaders) targeted for access and high-stakes actions

Older adults/low-digital-literacy contexts facing unsolicited calls/messages

Diaspora and remittance contexts – low-trust settings with limited verification capacity

Mechanism summary

Actors: Opportunistic fraudsters; organized scam rings; targeted harassers/coercers

Enabling conditions: Public samples; weak trust cues; low-friction, high-stakes workflows

Pathway and leverage points: Target → sample → clone → trust contact → trigger action; disrupt at recovery; admin change; payouts; trust cues

Likely adaptation: Multichannel laundering; account warming; threshold testing

Operating constraints (what shapes evidence and action)

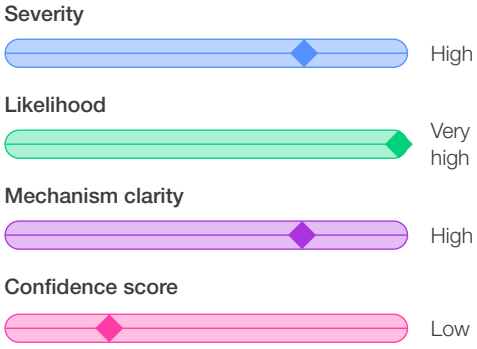
Visibility: Content/metadata/aggregates (often partial; proxies needed)

Relationship model: Direct users and/or tenant/admin-mediated workflows

Control surfaces: Trust cue; friction and throttles; enforcement; admin controls

External dependencies: Telcos/messaging routes; identity providers; payment rails; partners

Ratings (non-prescriptive)



Assess readiness snapshot

Coverage: Partial

Key gaps: Recovery/admin-change friction and step-up verification; detection leans on late signals and limited cross-channel visibility

Install early warning and assumptions

Early warning indicators:

- → External diffusion of kits/playbooks; sector incidents
- → Spikes in recovery attempts; admin/permission changes
- → New-device + sensitive action sequences; clustered reports/tickets

Assumptions and evidence:

Assumptions: Samples accessible; workflows engineerable at scale

Stronger: Confirmed incidents/probing/partner reports

Weaker: No precursor-impact linkage; mitigations reduce attempts without displacement

Routing to action pathways

Proposed pathway selection: Pathway 2 – Novel/uncertain coverage (mechanism plausible; coverage partial; confidence limited by visibility and adaptation)

Linked hypotheses: Fraud; coercion/harrassment; privilege compromise

Source: World Economic Forum

Not every possible harm uncovered in a foresight exercise is of equal threat value. Organizations should track the biggest and most time-sensitive risks. Without prioritization, organizations spread themselves thin, react to noise and miss compounding harms.

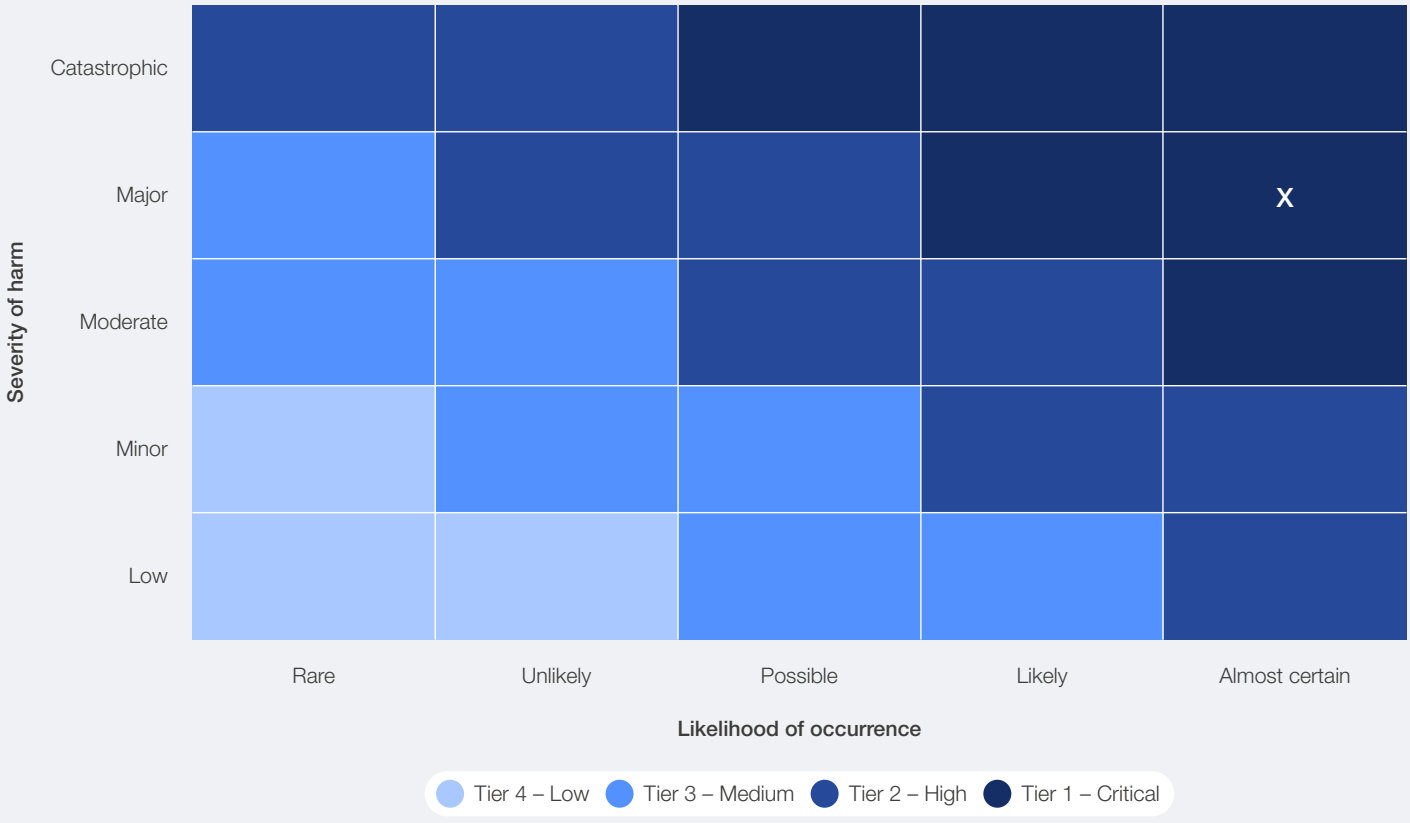
Teams should assess each scenario on at least two dimensions of risk:[34]

1. The severity of harm that can result from the considered hazard.
2. The probability of occurrence of that harm, which is a function of:
   - The exposure to the hazard
   - The occurrence of a hazardous event
   - The possibilities of avoiding or limiting the harm

Organizations can use qualitative or semi-quantitative scales for severity and likelihood and then combine them into tiers. These tiers should be linked to rules, such as funding, escalation and monitoring cadence so that prioritization drives action. Additionally, translating those levers into a score makes prioritization transparent and repeatable and helps leaders see which items to concentrate on.

Crucially, this severity–likelihood view is kept conceptually distinct from both the organization's clarity about the harm and its capacity to manage that harm effectively.

FIGURE 4 **Risk tiering heatmap (commodity AI impersonation kits)**



Source: World Economic Forum

2.7 Implementation in context

Foresight can be resource-intensive, and many trust and safety teams, especially in smaller organizations, do not have the personnel, tooling or time to run a full-scale programme. The intent of the pillars is therefore not to create an additional burden but to provide a structure that can be applied proportionately based on capacity, visibility and risk exposure. The guidance below outlines practical ways to use the same pillars more efficiently – by narrowing scope, standardizing lightweight documents, borrowing expertise selectively and using AI to reduce workload while keeping human accountability for judgement, escalation and decisions.

AI can reduce the cost of foresight work, but it is not a substitute for the operational infrastructure that turns signals into action. Every pillar depends on some underlying tooling and process layer. For smaller teams, that infrastructure is often the real constraint. Operationalizing the foresight pillars usually requires some reusable infrastructure. Shared and open safety infrastructure can help lower this barrier, especially where tools are policy-agnostic, modular and sufficiently flexible to be reconfigured as harm patterns evolve rather than being built from scratch.

Using AI

Where basic intake, tagging, review and escalation infrastructure already exists, AI is most immediately useful for horizon scans: here the workload is dominated by collecting, summarizing, translating, de-duplicating and tagging large volumes of information across diverse sources. Foresight practitioners already use AI predominantly for horizon scanning (60%) and trend analysis/clustering (69%), with AI framed as a supplement to human work rather than a replacement.[35] Once signals are collected, AI can help compress and silence ambient signals into a manageable set of drivers, proposing candidate clusters, labels and draft rationales. AI can both assist in capacity and highlight potential bias or blind spots.

AI can be used as a developer of plausible futures and scenarios through refinement and multiple variants that can then stress-test assumptions, identify contradictions and surface alternative pathways. The use of AI to generate these scenarios and then test them improves speed.

It can be difficult to make foresight a regular operating process. AI can help by drafting recurring outputs, maintaining traceable links from signals, drivers and triggers and supporting the monitoring of measurable triggers. The goal for employees is to focus their time and expertise on judgement and decisions.

> AI can reduce the cost of foresight work, but it is not a substitute for the operational infrastructure that turns signals into action.

AI integration should not be sudden. *AI in strategic foresight: Reshaping anticipatory governance*, an Organisation of Economic Co-operation and Development/World Economic Forum paper, presents a maturity path that can act as a ladder on the integration of AI:[36]

- **Level 1 – AI for analysis augmentation:** AI supports synthesis/scanning/sensemaking as standalone help and creates a base layer of insights that can then be analysed by experts.
- **Level 2 – AI as creative sparring partner:** AI helps generate ideas, systematize signals, suggest scenarios and stress test human content.
- **Level 3 – AI integrated and customized into workflow:** AI is integrated across the foresight process with tailored tools, including experimentation with agents to automate parts, such as signal detection, trend analysis, scenario development and stress testing – although this is rare.

Limiting strain and building capacity

Digital safety foresight can feel resource-intensive, especially for smaller teams in react mode. The pillars are meant to be used proportionately so organizations should narrow scope to between one and three priority harms at a time, do the work in a fixed time frame and rely on a small set of reusable documents, such as a one-page decision brief and short insight log. Used this way, foresight becomes a lightweight layer that strengthens everyday decision-making rather than a separate programme.

The aim is not comprehensive coverage; it is earlier, clearer judgement about what is changing, what is exposed and which actions are feasible. Where dedicated trust and safety tooling does not exist, teams should adapt existing operational systems, such as support, abuse, security or incident-management workflows rather than wait for a bespoke stack.

For Frame and govern, small teams should keep governance minimal, such as one scoping question, a defined horizon and explicit constraints. AI can reduce overheads by drafting, summarizing, de-duplicating and tagging sources – while a human owner remains accountable for final framing and escalation thresholds. Capacity builds when the decision brief is updated on a predictable cadence rather than rewritten, and when decision rights are clarified once so the team does not renegotiate authority during a crisis.

For Sense, it should be disciplined discovery, not constant monitoring. Maintain a limited watch portfolio that reflects the most relevant domains and likely spillover areas. AI is best used for collection and compression, such as with summarizing, clustering signals into themes and drafting short insight records. Humans should own interpretation of why it matters, confidence and constraint notes and the decisions about what moves forward. Capacity builds through consistent tagging (capabilities, incentives, access/distribution, evasion/adaptation) and periodic borrowing of expert review (short cross-functional or external check-ins) to reduce blind spots without adding permanent workload.

For Model, limited capacity should avoid heavyweight workshops and produce a minimum viable mechanism capture for each priority harm. This could include a short narrative that identifies leverage points, likely bypass categories and who is most affected. AI can accelerate this by turning inputs into generating alternative hypotheses, surfacing missing enabling conditions and suggesting likely adaptation routes – without publishing operational how-to details. Humans should sign off points that will drive readiness decisions. Through reuse, a small internal library of past one-pagers with patterns and common failure modes and a rotating facilitator will build capacity.

For Assess readiness, small teams should resist audit-style completeness and instead run a check focused on the most important points. AI can compress readiness work by summarizing past incidents and near misses, clustering recurring failure modes, drafting the control map from existing documentation and proposing candidate stress test scenarios. Humans must approve any readiness claim that implies capability and must review any action that affects access, enforcement or sensitive judgement. Capacity builds by tying readiness to existing operational structures, including incident response, customer escalation and security on call.

For Install early warning, organizations should monitor harms already prioritized and modelled, using a register with between three and eight indicators per harm. AI is useful for clustering tickets and reports and summarizing anomaly patterns. Again, a human owner should remain accountable. Through an indicator library containing accounts of what worked, a short review cadence and tight linkage to existing operational routines will prevent the invention of new processes.

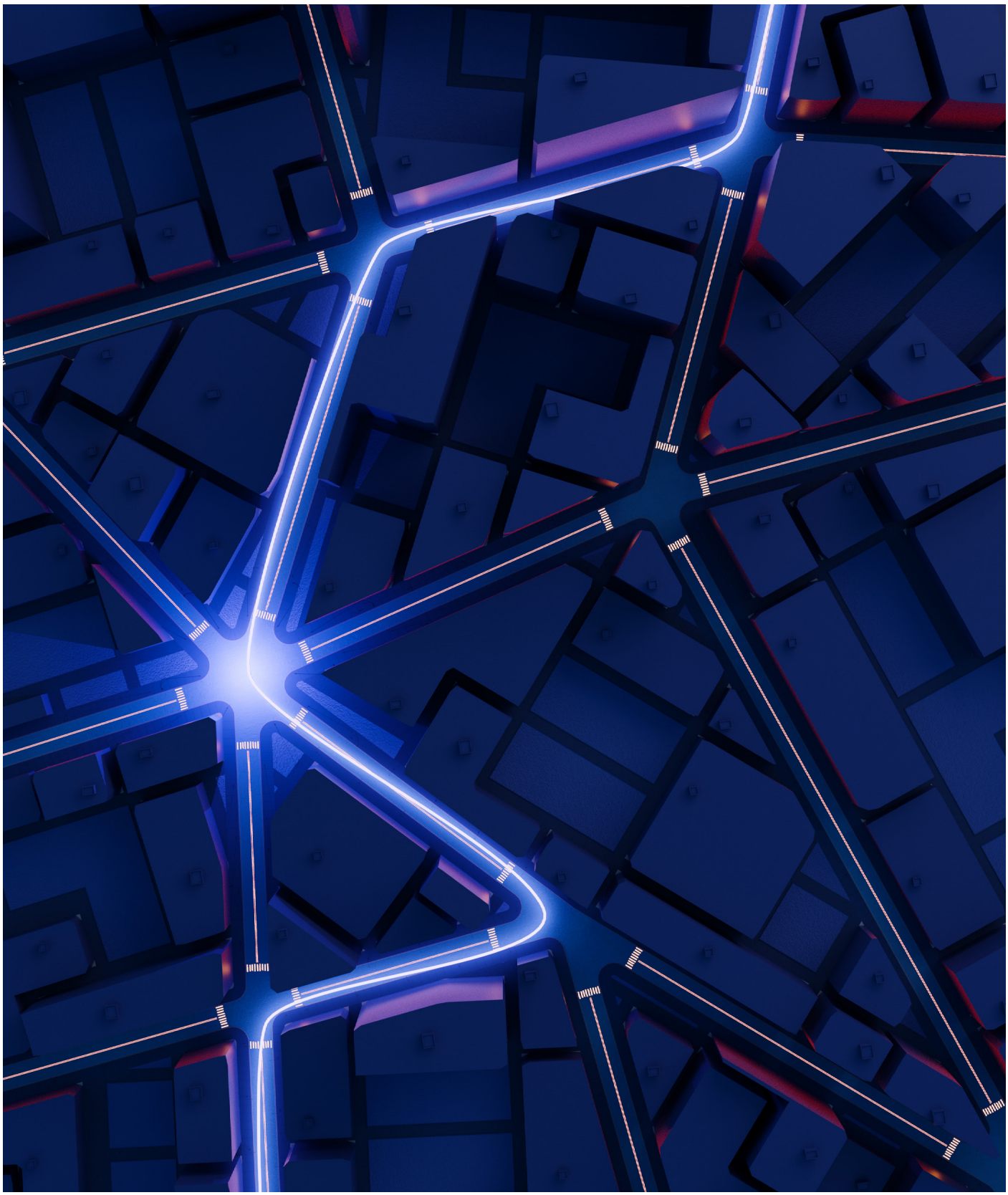
For Synthesize, AI can help assemble first drafts of the harm card from the underlying documents, but a human should review this, as the harm card will determine and justify organizational actions. Templates should be reused, and a record kept of prior harm cards and decisions.

Collaboration between organizations further reduces strain. Many services face the same emerging capabilities and adversary tactics at the same time; sharing general signals, using open-source tools, stress testing assumptions and comparing what has worked (or failed) reduces duplicated effort and improves judgement. Pooled learning is not outsourcing responsibility; it is a capacity multiplier that helps organizations.

A lightweight foresight practice is a form of operational resilience and risk management because it reduces costly surprise, shortens time-to-mitigation, improves cross-functional alignment and protects trust with users, customers and partners. It also strengthens accountability, as the organization can explain its decisions.

3 Response pathways

Four response pathways provide a mechanism for organizations to turn foresight into action.



The foresight process is valuable only if it leads to clear, timely decisions about what to do next. This section sets out response pathways that explain how organizations can move from identified future harms to modes of response:

1. Known and covered (enforce and tighten)
2. Novel/uncertain coverage (validate and prepare)
3. Not covered (targeted strengthening)
4. Unpredictable/unknown (tripwires and sentinels)

Instead of ad-hoc reactions, teams route each prioritized harm into one of these pathways in a consistent, explainable way. This routing is grounded in a combination of quantified scoring and qualitative justification produced by the previous foresight tool – in particular, how clearly the harm mechanism is understood and how confident the organization is that current protections can address it.

The actions in each pathway are not a list of required steps; they are examples of possible actions, to be adapted based on an organization's context, risk tolerance and capacity.

3.1 Pathway matrix

The matrix below provides four pathways that translate pillar outputs into a posture choice. Organizations should place a harm into one of the pathways using the synthesized harm card to make a qualitative judgement informed by quantified data. Placement should be anchored on mechanism maturity, coverage at key leverage points and operational constraints. Organizations may also apply modifiers that legitimately shift posture, such as severity/irreversibility, concentration on groups, time-to-scale or legal/regulatory exposure, so that being uncertain does not default to waiting. Each placement should be documented with a short rationale as well as evidence of what would justify moving the harm to a different pathway.

While the pathways provide a common structure for choosing a response posture, the specific actions an organization takes will necessarily vary based on its service type, operating model, visibility constraints and available control surfaces. For example, a consumer platform may respond through ranking, moderation or in-product reporting, while a cloud or enterprise provider may act through suspicious-activity detection, account or tenant controls, customer notification, acceptable-use enforcement or takedown and escalation processes. Organizations should therefore treat each pathway as a direction of travel to maintain, validate, strengthen or build capacity, then tailor the concrete interventions to what is feasible and relevant.

TABLE 2 **Response pathway matrix**

PATHWAY 1 **Known and covered** (enforce and tighten)	The harm mechanism is well understood; recent evidence shows strong coverage at the key leverage points and the current threat environment has not materially changed the assumptions that support this coverage. The posture is to calibrate existing controls, expand them where needed and watch for drift, gaming and capability shifts that would require revalidation. Keep monitoring and escalation routes current.
PATHWAY 2 **Novel/uncertain coverage** (validate and prepare)	The harm mechanism is emerging or evidence is directional but there is a credible hypothesis for how the harm could materialize in the context and it may be partially covered or coverable with low-regret changes. The posture is to validate the most decision-critical assumptions, improve observability or proxies where visibility is limited, stand up a lightweight indicator set tied to leverage points and pre-position mitigations that can be activated without waiting for full certainty.
PATHWAY 3 **Not covered** (targeted strengthening)	The harm mechanism is well understood, but the readiness work shows thin or clearly insufficient coverage at the leverage points that determine impact. The posture is to build or strengthen specific barriers, close known failure modes, clarify decision rights and execution pathways and reduce reliance on assumptions that the system will hold under adversarial adaptation.
PATHWAY 4 **Unpredictable/ unknown** (tripwires and sentinels)	The harm mechanism is poorly understood and/or the constraints make both detection and mitigation uncertain, and current protections are untested or unlikely to be sufficient. The posture is to establish minimum viable intake and escalation, create proxy observables, determine pre-agreed tripwires and escalation authority, define redress and communications paths and build the basic response infrastructure needed so that, as evidence accumulates, the issue can be moved into validation/preparation or strengthening rather than remaining unactionable.

PATHWAY 1

Known and covered

Pathway 1 is chosen when the harm is well understood and current protections are judged to be broadly effective. The problem is not only consistency, coverage and drift but whether the current evidence still justifies confidence in those controls under changed conditions.

This pathway should be viewed less as innovation and more as reliability engineering for safety. The rationale is to prevent complacency. If a harm is known and covered, the organization is accountable when it resurfaces due to regressions, uneven enforcement or blind spots for exposed groups. Placement in Pathway 1 should clear a higher bar than familiarity; it should be supported by evidence that the harm mechanism is still materially the same, that coverage exists at the key leverage points and that recent audits, incident patterns, evaluations, red-team findings or front-line operational reviews show that those controls still perform under current conditions, including for highly exposed groups and edge contexts.

Once a harm is determined to be on Pathway 1, organizations should treat it as a disciplined tightening exercise:

- **Revalidate coverage assumptions:** Use targeted audits and sampling to check whether existing policies, models, product features and human workflows are in fact intercepting the harm. Pay particular attention to high VEC groups and lower-resourced languages or regions.
- **Standardize and strengthen enforcement:** Where audits reveal inconsistency, invest in clearer decision guides, training and calibration sessions. For automated systems, revisit thresholds and tuning to ensure that the harm is not slipping through in pursuit of other objectives.
- **Guard against regressions:** Make the harm part of change management by building regression tests into model updates and product launches; add checks to launch gates so that new features cannot ship without confirming that controls for this harm still function. Organizations may seek to define explicit levels of safety service levels. This may look like minimum detection rates or maximum tolerated incident levels that must be maintained over time.
- **Improve user-facing tools and redress:** Even where controls are strong, affected users may struggle to find or use them. Safety information should thus be communicated urgently in accessible, trusted formats. In this pathway, refinements that reduce friction might include clearer reporting categories, better in-product explanations, faster and more transparent appeals and support for those affected.
- **Share learning internally and, where possible, externally:** Because these harms are already known, there is an opportunity to codify what works into playbooks, case studies and internal safety patterns that other teams can reuse. Where feasible, sharing high-level lessons with peers or civil-society partners can help raise the baseline across the ecosystem.

Pathway 1 should not become a parking lot for issues that feel familiar. It is a commitment to keep known harms under control, especially for those most affected, and to treat safety regressions with the same seriousness as other major quality failures. Placements should also be reopened when the landscape shifts materially, even if policies and tooling appear unchanged. Capability jumps can make an old harm operationally different.

PATHWAY 2

Novel/uncertain coverage

Pathway 2 applies when the harm is not yet fully understood, but it is plausible that existing protections might be sufficient or could be extended with adjustments. Here, the organization does not want to overreact by building entirely new systems on the basis of speculation, nor underreact by assuming that everything is fine. The rationale is to learn quickly under protection- and design-focused tests that reduce uncertainty about mechanisms and coverage, while limiting additional exposure.

Once a harm is placed into Pathway 2, the work is structured around time-bound learning:

- **Guard against regressions:** Move beyond questions of danger to questions that can be tested, such as "Do current classifiers detect at least X% of this pattern?", "Are Group A seeing materially higher incidence than others?" or "Does this new behaviour bypass our current limits?" These questions anchor experiments and monitoring.
- **Design safe experiments and probes:** Depending on the harm, this might include targeted red teaming, simulations, sandboxed deployments, constrained A/B tests or synthetic data exercises. Experiments should be pre-approved through a governance process that weighs the potential insight against any incremental risk and should always include a way to halt or roll back if concerning effects appear.
- **Add temporary guardrails while learning:** While tests run, it is prudent to include conservative friction, such as tighter rate limits in high-risk contexts, stronger defaults or additional review for certain actions. This should be implemented especially with high-severity impacts. Such temporary mitigations buy time to learn without committing to long-term changes.
- **Monitor both internal and external signals:** Early-warning indicators previously defined should be tracked at a higher cadence and maintain open lines with external partners who may see early cases. Divergence between internal metrics and external reports should be treated as a signal that coverage assumptions may be wrong.
- **Decide and reroute:** At the outset, set clear timebox and decision criteria. Determine what specific findings would justify a change in pathways. At the end of the period, update the harm profile as appropriate and formally reroute.

Pathway 2 functions as a safety research mode and allows organizations to probe emerging harms, refine their understanding and avoid both complacency and overbuilding, all while keeping users as protected as possible during the learning process.

PATHWAY 3

Not covered

Pathway 3 is chosen when the harm is understood, but current protections do not substantially cover it. The organization can see how, where and to whom the harm would occur, and the control map shows that existing policies, tools and processes are absent or misaligned. If a serious harm is likely and existing measures are insufficient to prevent or mitigate it, new safeguards must be designed and implemented before exposure expands.

Once a harm is routed onto Pathway 3, the focus shifts to deliberate design and build:

- **Define a clear protection brief:** Start from the harm profile and control map to write a brief on which mechanisms need intervention, which groups and contexts must be protected, what constraints apply (legal, technical, business) and what success would look like. The brief should explicitly reference severity, reversibility and equity considerations.
- **Co-design interventions with affected groups:** Where feasible, involve representatives of high VEC groups, front-line staff and relevant external experts in shaping solutions. They can highlight failure modes, such as discouraged usage, misuse of reporting tools or unintended exclusion that internal teams might miss, and help ensure that new controls do not shift burdens onto those already at risk.
- **Build a layered intervention set:** New protections rarely consist of a single tool. For each harm, consider a combination of:
  - Product and user experience (UX) changes that remove or reduce risky features or add friction in high-risk contexts
  - New or upgraded detection and ranking systems

- Strengthened policies and enforcement guidelines
- Enhanced user education, reporting and redress
- Operational processes (for example, special processes for unusual or sensitive cases)

The interventions should be proportional and strong enough to meaningfully reduce harm, but should not unnecessarily restrict legitimate use.

- **Gate scale behind safety readiness:** Major launches, expansions to new markets or rollouts to high-risk populations should not proceed until the agreed minimum safeguards for the harm are in place and have passed initial evaluation.
- **Evaluate and iterate:** Use a mix of offline evaluation, live studies and qualitative research to test whether new interventions are working. Monitor both direct harm indicators and second-order impacts, such as false positives, user confusion or shifts in adversary behaviour. Based on results, refine interventions and update the harm profile.
- **Strengthen redress alongside prevention:** Even with new safeguards, some harm will occur. Ensure that affected users have accessible routes to complain, appeal and receive support, and that severe cases can be escalated quickly. In Pathway 3, building new protections should go hand in hand with building more responsive and fair redress.

Pathway 3 is resource-intensive by design. The organization is choosing to invest in new safety infrastructure because it has sufficient clarity about the harm and enough knowledge about current gaps.

PATHWAY 4

Unpredictable/unknown

Pathway 4 is reserved for harms that are poorly understood and where existing protections are unlikely to be sufficient or are largely untested. There are harms the organization cannot yet describe clearly, let alone control, and for which it is not prepared. That is a practical problem with designing responses when neither the risk nor the remedy is well specified. It asks leaders and teams to acknowledge uncertainty while still taking responsibility and acting cautiously, listening and building protections in the dark.

The rationale here is not to pretend to know the unknowable, but to build resilience amid uncertainty by putting in place tripwires that catch concerning patterns, and sentinels that can interpret signals as well as robust redress. This pathway reflects an inherent limit to being fully prepared, but it should not lead to constant scenario planning.

This requires more than monitoring dashboards. It calls for a distinct posture built around humility, reversibility and preparedness.

Designing tripwires

Tripwires are pre-agreed conditions that, if met, trigger a response, including pausing a feature, tightening controls or convening a crisis team. As in the foresight pillars, this aspect requires clear authority to act and clarity on required documentation.

Pathway 4 is where organizations are most likely to face pressure not to act. The costs of precaution (slowed features, added friction, delayed launches) are immediate and visible. The costs of inaction are delayed and often borne by others. This means that waiting for more evidence tends to win by default,

even as signals accumulate. Pre-agreed tripwires help counter this by anchoring the decision to act in a commitment made before the pressure of the moment, so that precautionary steps do not depend on individuals advocating against short-term costs in real time

Due to the nature of Pathway 4, rather than seeking a single, well-defined signal, the goal is to identify patterns that indicate when something is wrong, even if the specific issue is not yet known. Tripwires should be documented in advance, linked to the early-warning indicators and designed to notice what is usually not looked at, not just spikes in the usual metrics.

Useful tripwires include:

- **Leading indicators of stress or lack of stress:** For example, sudden spikes in unusual clusters of appeals, abnormal usage patterns in sensitive contexts or correlated anomalies across products. But in the unknown case, they can also be inverted patterns, such as a sudden drop in use by an active user. These indicators will not be perfectly specific, but they can flag that something is awry earlier than fully specified metrics.

- **Thresholds on harm proxies:** Where direct measurement is difficult, organizations can choose conservative proxies – for instance, a maximum tolerated rate of serious complaints in high-risk regions or bounds on how much a new capability can be used in contexts (elections, conflicts, civic processes). Crossing those bounds does not prove a new harm, but it triggers review and, if needed, activates throttling or targeted guardrails until the situation is better understood.

- **Circuit-breakers for new behaviours:** For features or models with high systemic potential, such as generative tools, automation surfaces or cross-platform integration the conditions under which usage will be slowed, narrowed or paused should be predefined. This can include credible evidence from partners of emerging abuse patterns or unexplained escalations in risk signals.

Building sentinels

Sentinels are the people and structures that notice, interpret and elevate signals that do not yet fit into tidy metrics. Organizations should:

- **Establish internal sentinel roles:** This might include a cross-functional group drawn from trust and safety, policy, product, operations and data science, with a mandate to review information relating to risks. They should have direct lines of communication to senior decision-makers and the authority to recommend precautionary action.

- **Invest in external sentinel networks:** Build and maintain relationships with communities, nongovernmental organizations (NGOs), researchers and journalists who may see emerging harms first. Just as in the foresight pillars, provide them with safe, structured channels to report concerns; offer, where appropriate, support, such as contacts, context or data access; be clear about how their inputs will be used.

- **Enable internal escalation pathways:** Encourage employees, especially those in front-line roles, to flag unusual harms without fear of blame or retaliation. Simple internal reporting tools and regular sentinel reviews of such reports can surface early warnings.

Redress when the unknown materializes

In all pathways, but especially this one, organizations must assume that some harms will bypass their foresight. Redress is therefore central, not secondary. What is needed is:

- **Accessible, flexible complaint routes:** Make it easy for individuals and organizations to say when something did not go as planned. Open-ended descriptions and support from human reviewers are important to complement structured forms.

- **Fair and timely handling of novel cases:** Establish teams that can handle unfamiliar harms, consult subject-matter experts and avoid dismissing reports simply because they do not match existing taxonomies. Where users have suffered significant harm, consider remedies beyond standard account actions, including outreach, tailored product changes or referrals to support services.

- **Transparent acknowledgement and learning:** When a previously unforeseen harm is confirmed, the organization should acknowledge it, at least in aggregate terms, and explain what is being done in response. Internally, this should trigger an after-action review that feeds back into foresight, including rerouting the harm into Pathways 1–3 as it becomes better understood and controlled.

Pathway 4 is demanding because it asks organizations to operate in uncertainty, build modest but real safeguards against escalation and prioritize redress when things go wrong.

Conclusion

Digital safety harms will not stand still. As AI capabilities mature, synthetic identity becomes easier to deploy and adversaries adapt across services, organizations face a common challenge. The conditions that enable harm are changing fast. It is not enough to rely only on retrospective enforcement or product-specific risk review. Organizations need a structured way to look outward, detect change early, test their assumptions and translate uncertainty into practical decisions. Providing such a structure is the central contribution of this report.

The digital safety foresight pillars are intended to support organizations with the key elements and practices not only to predict the future but to better prepare themselves for it. Effective foresight translates uncertainty into concrete decisions. The pillars set out in this report – Frame and govern, Sense, Model, Assess readiness, Install early warning and Synthesize – are designed to make that translation possible in a structured way.

The four response pathways make this discipline operational beyond understanding the risks. Not all organizations will face every emerging harm with the same degree of clarity or preparedness. Some harms are already known and require stronger consistency, enforcement and maintenance of protections. Some are novel but may be addressed through focused testing and adaptation. Others show that current safeguards are inadequate and that new controls must be built. Still others remain uncertain and poorly specified, requiring organizations to act cautiously with tripwires, sentinel networks and robust redress while knowledge is still incomplete. The goal of foresight is not to eliminate uncertainty but to prevent uncertainty from becoming an excuse for drift, delay or avoidable harm.

A key message throughout this report is that effective foresight must be proportionate and usable in the real world. It cannot be reserved only for the best-resourced organizations or treated as a separate strategic exercise detached from daily operations. Less-resourced teams can apply the same logic through narrower scopes, lightweight documents, selective collaboration and bounded uses of AI to support them while retaining human judgement and accountability.

Some harms will still emerge through blind spots, spillovers and new combinations. For that reason, foresight must be paired with humility, strong feedback loops and meaningful redress. What matters is not whether an organization can predict every future risk perfectly but whether it can build the institutional habits to learn early, respond proportionately, explain its choices and improve over time so they are not caught on the back foot. Organizations that do this well will be better able to protect users, build trust and adapt responsibly to the evolving digital environment. Ultimately, digital safety foresight should be understood not as a one-off exercise but as a core governance capability.

Contributors

Lead author

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

World Economic Forum

Agustina Callegari
Initiatives Lead, Technology Governance, Safety and International Cooperation

Cathy Li
Head, Centre for AI Excellence; Member of the Executive Committee

Acknowledgements

This report 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 listed below.

Sincere appreciation is extended to the following members and other experts, 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

Daniel Child
Manager, Industry Insights and Enablement, Office of the eSafety Commissioner, Australia

Jeffrey Collins
Director, Trust and Safety, Amazon Web Services

Ben Colman
Chief Executive Officer, Reality Defender

Kieran Donovan
Chief Executive Officer and Co-Founder, k-ID

Iain Drennan
Executive Director, WeProtect Global Alliance

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

Courtney Gregoire
Chief Digital Safety Officer, Microsoft

David Evan Harris
Chancellor's Public Scholar, UC Berkeley

Sasha Havlicek
Chief Executive Officer, Institute for Strategic Dialogue

Lisa Hayes
Head, Safety Public Policy; Senior Counsel, TikTok

Adeline Hulin
Head of Unit for Media and Information Literacy and Digital Competencies, United Nations Educational, Scientific and Cultural Organization (UNESCO)

Lea Kasper
Executive Director, Global Partners Digital

Renato Leite Monterio
Vice-President, Privacy, Data Protection, AI and Intellectual Property, e&

Sunny Xun Liu
Director of Research, Stanford Social Media Lab

Marija Manojlovic
Director, Safe Online

Mauro Miedico
Director, United Nations Counter-Terrorism Centre

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

Danielle Osler
Director, Trust and Safety Global Engagement, Google

Juliet Shen
Head of Product, Robust Open Online Safety Tools (ROOST)

Rebecca Smith
Global Head of Child Protection Programmes, Save the Children International

Ian Stevenson
Chair, Online Safety Tech Industry Association (OSTIA)

David Sullivan
Executive Director, Digital Trust and Safety Partnership

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

Hayley van Loon
Chief Executive Officer, Crime Stoppers International

David Wright
Chief Executive Officer, South West Grid for Learning (SWGfL)

Production

Laurence Denmark
Creative Director, Studio Miko

Charlotte Ivany
Designer, Studio Miko

Simon Smith
Editor, Astra Content

World Economic Forum Public Engagement

Maxwell Hall
Creative Editorial Lead

Floris Landi
Design Lead

Gayle Markovitz
Head, Written and Audio Content

Sybile Penhirin
Head, Video and Design

Endnotes

1. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*, p. 2. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf
2. Faverio, M., & Sidoti, O. (2025, December 9). *Teens, social media and AI chatbots 2025*. Pew Research Center. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/
3. Hoffner, C. A., & Bond, B. J. (2022). Parasocial relationships, social media, & well-being. *Current Opinion in Psychology, 45*, article 101306. https://pubmed.ncbi.nlm.nih.gov/35219157/
4. Anthropic. (2024, October 22). *Introducing computer use, a new Claude 3.5 sonnet, and Claude 3.5 haiku*. https://www.anthropic.com/news/3-5-models-and-computer-use; OpenAI. (2025, January 23). *Introducing Operator*. https://openai.com/index/introducing-operator/
5. Schroeder, D. T., et al. (2026). How malicious AI swarms can threaten democracy. *Science, 391*(6783), 354–357. https://pubmed.ncbi.nlm.nih.gov/41570131/
6. Piscitello, D. (2025, September 9). *Phishing landscape 2025: An annual study of the scope and distribution of phishing*. Interisle Consulting Group. https://interisle.net/insights/phishing-landscape-2025-an-annual-study-of-the-scope-and-distribution-of-phishing
7. Bellan, R. (2025, December 22). *OpenAI says AI browsers may always be vulnerable to prompt injection attacks*. TechCrunch. https://techcrunch.com/2025/12/22/openai-says-ai-browsers-may-always-be-vulnerable-to-prompt-injection-attacks/
8. Simon, F., Nielsen, R. K., & Fletcher, R. (2025). *Generative AI and news report 2025: How people think about AI's role in journalism and society*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society
9. Financial Crimes Enforcement Network. (2024, November 13). *FinCEN alert on fraud schemes involving deepfake media targeting financial institutions*. https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf
10. Leng, C., & Chan, H. (2024, May 16). Arup lost $25mn in Hong Kong deepfake video conference scam. *Financial Times*. https://www.ft.com/content/b977e8d4-664c-4ae4-8a8e-eb93bdf785ea
11. Kira, B. (2024). When non-consensual intimate deepfakes go viral: The insufficiency of the UK Online Safety Act. *Computer Law & Security Review, 54*, article 106024. https://www.sciencedirect.com/science/article/pii/S0267364924000906; UN Women. (2025). *Tipping point: The chilling escalation of online violence against women in the public sphere in the age of AI*. https://www.unwomen.org/en/digital-library/publications/2025/12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai
12. Timoney, M. (2025, 17 April). *Gen AI is ramping up the threat of synthetic identity fraud*. Federal Reserve Bank of Boston. https://www.bostonfed.org/news-and-events/news/2025/04/synthetic-identity-fraud-financial-fraud-expanding-because-of-generative-artificial-intelligence.aspx
13. Shumailov, I., et al. (2023). *The curse of recursion: Training on generated data makes models forget*. arXiv. https://arxiv.org/abs/2305.17493; World Economic Forum. (2025). *Synthetic data: The new data frontier*, p. 8. https://reports.weforum.org/docs/WEF_Synthetic_Data_2025.pdf
14. Coalition for Content Provenance and Authenticity. (2024). *Content credentials: C2PA technical specification*. https://spec.c2pa.org/specifications/specifications/2.1/specs/_attachments/C2PA_Specification.pdf
15. TikTok. (2024, May 9). *Partnering with our industry to advance AI transparency and literacy*. https://newsroom.tiktok.com/partnering-with-our-industry-to-advance-ai-transparency-and-literacy?lang=en
16. Coalition for Content Provenance and Authenticity. (2022). *C2PA security considerations*. https://spec.c2pa.org/specifications/specifications/1.0/security/Security_Considerations.html
17. McClain, C., & Bishop, W. (2026, January 8). *What we know about internet use, smartphone ownership and digital divides in the U.S.* Pew Research Center. https://www.pewresearch.org/short-reads/2026/01/08/internet-use-smartphone-ownership-digital-divides-in-u-s/; Gelles-Watnick, R. (2024, January 31). *Americans' use of mobile technology and home broadband*. Pew Research Center. https://www.pewresearch.org/internet/2024/01/31/americans-use-of-mobile-technology-and-home-broadband/
18. United Nations Educational, Scientific and Cultural Organization (UNESCO). (2025). *Recommendation on the ethics of neurotechnology*. https://www.unesco.org/en/ethics-neurotech/recommendation
19. Ruder, K. (2025, July 23). *States pass privacy laws to protect brain data collected by devices*. KFF Health News. https://kffhealthnews.org/news/article/colorado-california-montana-states-neural-data-privacy-laws-neurorights/
20. MacInnes, P. (2025, September 15). Amazon to offer Champions League viewers new immersive in-game data. *The Guardian*. https://www.theguardian.com/football/2025/sep/15/amazon-prime-video-prime-vision-data-champions-league-football; Sprigg, S. (2025, July 21). *Kinneta introduces AR and VR workouts for fitness machines*. Auganix. https://www.auganix.org/vr-news-kinneta-xr-workouts/; XR Today. (2026, March 25). *XR market expands 44.4% in 2025 as smart glasses take center stage*. https://www.idc.com/promo/arvr/

21. Mehta, I. (2025, May 1). *WhatsApp now has more than 3 billion users a month*. TechCrunch. https://techcrunch.com/2025/05/01/whatsapp-now-has-more-than-3-billion-users/

22. Udupa, S., et al. (2025, September 5). *Extreme speech in encrypted messaging: Recommendations for holistic policy*. MediaWell. https://mediawell.ssrc.org/articles/extreme-speech-in-encrypted-messaging-recommendations-for-holistic-policy/; National Society for the Prevention of Cruelty to Children (NSPCC). (2025). *Data shows how criminals are using private messaging platforms to manipulate and groom children*. https://www.nspcc.org.uk/about-us/news-opinion/2025/data-shows-how-criminals-are-using-private-messaging-platforms-to-manipulate-and-groom-children/; Olaizola Rosenblat, M., Trauthig, I. K., & Woolley, S. C. (2024). *Covert campaigns: Safeguarding encrypted messaging platforms from voter manipulation*. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/wp-content/uploads/2024/10/NYU-CBHR-Covert-Campaigns_FINAL-FINAL-Sep29.pdf

23. Graphika. (2025, August 20). *Jumping off platform*. https://www.graphika.com/reports/jumping-off-platform

24. NSPCC. (2025). *Data shows how criminals are using private messaging platforms to manipulate and groom children*. https://www.nspcc.org.uk/about-us/news-opinion/2025/data-shows-how-criminals-are-using-private-messaging-platforms-to-manipulate-and-groom-children/

25. European Commission. (n.d.). *The Digital Services Act*. Retrieved April 10, 2026, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

26. EU Artificial Intelligence Act. (2024). *The AI Act explorer*. https://artificialintelligenceact.eu/ai-act-explorer/

27. UK Parliament. (2023). *Online Safety Act 2023*. https://www.legislation.gov.uk/ukpga/2023/50/contents

28. Australian Government eSafety Commissioner. (2025). *Social media age restrictions*. https://www.esafety.gov.au/about-us/industry-regulation/social-media-age-restrictions

29. Infocomm Media Development Authority. (2025, May 14). *Enhancing online safety in Singapore*. https://www.imda.gov.sg/regulations-and-licensing-listing/content-standards-and-classification/standards-and-classification/internet/online-safety

30. Global Online Safety Regulators Network. (2024, April). *Regulatory coherence and coordination: The role of the Global Online Safety Regulators Network* [Position statement]. https://www.esafety.gov.au/sites/default/files/2024-05/GOSRN-Position-Statement-on-Regulatory-Coherence.pdf

31. Olaizola Rosenblat, M., Agrawal, A., & Yap, I. (2025). *Online safety regulations around the world: The state of play and the way forward – A resource guide*. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/wp-content/uploads/2025/04/NYU-CBHR-Online-Regulations_Updated-Jun-17-1.pdf

32. Casalini, F., López González, J., & Nemoto, T. (2021). *Mapping commonalities in regulatory approaches to cross-border data transfers*. Organisation for Economic Co-operation and Development (OECD) Trade Policy Papers No. 248. OECD Publishing. https://www.oecd.org/en/publications/mapping-commonalities-in-regulatory-approaches-to-cross-border-data-transfers_ca9f974e-en.html

33. Internet Society. (2025). *Age restrictions and online safety*. https://www.internetsociety.org/resources/policybriefs/2025/age-restrictions-and-online-safety/

34. International Organization for Standardization and International Electrotechnical Commission (ISO/IEC). (2014). *ISO/IEC Guide 51:2014: Safety aspects – guidelines for their inclusion in standards*, p. 3. ISO. https://www.iso.org/resources/publicly-available-resources.html?t=Ruvyk3OTE1FejG9wJie0LojrUDDf3weEqbMekT5NvH7ARf5jR7ng2dLlMyUmadiO&view=documents#section-isodocuments-top

35. World Economic Forum. (2025). *AI in strategic foresight: Reshaping anticipatory governance*, p. 8. https://www.weforum.org/publications/ai-in-strategic-foresight-reshaping-anticipatory-governance/

36. Ibid., p. 9.

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