SHIELD Global Online Safety Conference

REFERENCE DOCUMENT 2026

Insights, Findings & Global Voices

80 Speakers | 25 Countries

Shield.

How to Read This Document

This document brings together insights, arguments, and lived realities from practitioners, young people, researchers, and community leaders who spent three days mapping where digital harm originates and what it actually takes to address it. Some contributors appear without being named individually due to safeguarding practices.

To navigate it:

- **Start with the Four Learnings** for a distilled account of what surfaced across three days, regions, harm types, and disciplines.
- **Read the Voices sections** to understand how practitioners experience online harm in real time and what they're building to respond to it.
- **Consult the Session Reference** to locate specific speakers and arguments. The indexes trace recurring themes: resilience, co-design, identity systems, AI governance, disclosure, and the structural conditions that connect them.

Each section stands alone, but certain patterns recur across regions and disciplines: online harm originates below where responses land, populations are excluded from the design of systems meant to protect them, and practitioners build past that gap with little institutional recognition.

Note on Sources

Factual claims and references in this document are cited using a numbered in-text reference system. Sources are indicated in the text with bracketed numbers (e.g., [1]) corresponding to a numbered reference list provided at the end of the document.

About SHIELD

SHIELD – APS is a nonprofit organization headquartered in Lecce, Italy, dedicated to strengthening oline safety, protecting children and vulnerable communities, and advancing responsible technology worldwide.

Our mission is to expand collective capacity to understand and respond to how the online world is impacting individuals and communities by elevating lived experience, supporting community-led innovation, and creating environments where safety, dignity, and human rights are central to how technology is designed, governed, and used. SHIELD convenes international conferences, builds cross-sector coalitions, supports youth-led initiatives, and partners with organizations across the globe to strengthen local ecosystems of safety.

Legal Form: Associazione di Promozione Sociale (APS)
Registration Number (Fiscal Code): 93174300751
Operational Email: contact@shieldthefuture.com
Media & Partnerships: contact@shieldthefuture.com
Website: https://shieldthefuture.com

SHIELD works through a global aliance network and distributed model that centers regional leadership, intergenerational collaboration, and the belief that the people closest to online harm hold essential knowledge for building safer digital futures.

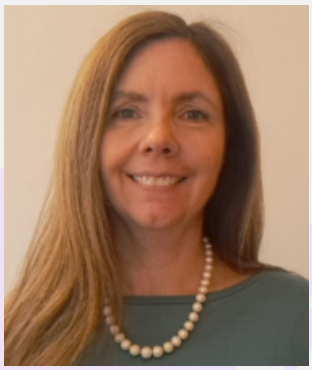
A Message from the Executive Director

In March 2026, eighty speakers and contributors from twenty-five different countries spent three days answering two questions: why does online harm persist, and what does it actually take to address it? The response is structural. The people closest to harm are building solutions that work: disclosure systems in languages formal platforms don't monitor, resilience frameworks designed with communities rather than for them, safety tools that don't demand surveillance as the price of protection. But they are doing so without institutional recognition, without adequate funding, and often without knowing others are solving parallel problems elsewhere.

When practitioners working at ground level are brought into the same space, patterns emerge. The harms technology produces are global, but the solutions are almost never universal. They are local, built from the ground up by people working inside the conditions they are addressing. The connection and awareness of these efforts matter. When practitioners can see patterns across contexts, they recognize what might transfer to their own work. But what is needed is not the unification of efforts under a single approach. It is the equalization of resources so that local solutions can draw from a global network on their own terms.

Online safety currently operates on two theories of change. One works through policy frameworks, platform accountability, and institutional reform. The other works at the grassroots level, driven by practitioners responding directly to the people and communities they serve. Both matter equally. But the resources, recognition, and infrastructure have concentrated almost entirely in the first, while the second produces solutions that actually function where harm occurs.

This document makes visible what has been building at ground level. The practitioners gathered here have already proven that their approaches work in their local realities. What happens next depends on whether the resources, recognition, and infrastructure follow the evidence.



Angeline Corvaglia
Executive Director, SHIELD - APS

TABLE OF CONTENTS

Conference Summary

Voices of the Conference

Session Reference

Indexes

Where the Conference Was Sitting

This section outlines the conditions that shaped the conversations across the three days. It situates the conference within the wider landscape of digital safety work, where long-established institutional approaches operate alongside emerging community-based practices. Understanding this context helps clarify why certain efforts gain traction, why others struggle to be recognized, and why some of the most effective insights come from people working closest to harm.

Online safety work is currently organized around two distinct theories of change that rarely engage with one another directly.

Theory of Change #1: Institutional Leverage

The first theory of change, developed over decades by large NGOs and academic institutions, is built around leverage points like legislation and platform accountability. This approach has produced essential, hard-won wins; major platform regulation exists today specifically because people in this field pushed back against well-funded resistance.

But it has a structural blind spot, and the conference repeatedly exposed it. It requires harm to be legible to institutions for them to act on it. Legibility, which is the ability of a harm to be categorized and measured by an institution, is not the same as visibility. A harm can be completely visible to the people experiencing it and still be illegible to institutional actors because it is not quantified properly, does not map onto existing legal categories, is too geographically or culturally specific to generalize, or involves communities that were never part of the design conversation.

Large institutions operate like a wide-angle lens, built to see the big picture. But that lens has a structural limit: it can only 'focus' on harms that are quantified or categorized. This means practitioner knowledge, which is often highly specific and 'high-resolution,' remains invisible not by choice, but because of the hardware of the system. Ultimately, this invisibility is a feature of institutional legitimacy; to remain a 'valid' authority, the system must prioritize the data it can process over the messy reality it cannot yet name.

Theory of Change #2: Ground-Level Practice

The second theory of change operates at the grassroots level. It is driven by practitioners who work in the context of their local needs, building local-language platforms and privacy-first tools for communities that distrust centralized data collection. They define harm through direct, lived experience in the environments where it actually occurs.

These practitioners build from the varied needs of specific people in specific places. That specificity is not a weakness. It is the key to effectiveness. While a universal framework provides the necessary infrastructure for global change, ground-level work captures the nuances that a broader lens necessarily overlooks. This approach recognizes that a tool built without the community it serves will struggle to protect that community. A tool built with the community it serves provides a layer of protection that anything created at a distance cannot.

This approach maps reality as it is found in the languages communities speak and within the trust relationships they have built. Because digital harm is not a single thing with a single cause, it cannot have a single solution. The practitioners who understand this most clearly are often the ones furthest from the rooms where decisions get made; they are the keepers of the granular data that completes the picture of safety that the larger system is built to manage.

Key Learnings: Four learnings emerged consistently across all three days of the conference, across regions, speaker backgrounds, and harm types. They are presented here as learnings rather than conclusions because the evidence supporting them is real, and the implications remain contested. Each learning is described as the conference presented it.

Learning 1:
Evidence of Safety Is Not the Same as Safety

Safety systems that measure process rather than outcome risk producing evidence of safety without producing safety itself.

The Design Problem

Evidence of safety is what a system can point to: reports filed, content removed, policies published, audits passed. It is measurable, documentable, and legible. What it does not measure is whether the person the system was built for is actually safer. A system oriented toward producing that evidence serves the organisation that needs to demonstrate compliance, not the person who needs to be protected. Remedy needs to be built into the governance architecture as a required outcome, not left to individuals to pursue through litigation after the system has already failed them.

The Accountability Gap

Governance structures designed to prevent online harm often serve more as records of accountability than as mechanisms to ensure it. Multiple layers of oversight, platform policy, terms of service, and regulatory frameworks can each show compliance, yet together they often leave the person harmed with no clear way to seek remedy. The presence of a safety mechanism and its ability to actually protect are not the same thing. For instance, when accountability is spread across layers that do not communicate, no single layer is responsible for the results.

This pattern appears across harm types and geographies. Parental control tools that report message volumes but tell parents nothing about risk. Disclosure systems that respond to a child's report by confiscating the device. Platform moderation frameworks that can demonstrate compliance with their own standards while the harm they were designed to address continues. In each case the system is functioning as designed. The design, however, is not oriented toward the person it was built to protect.

The Metric Trap

When safety is defined by internal metrics such as response times or volume of content removed, the system optimizes for those numbers rather than the reduction of harm. This creates a "success" state on paper that coexists with failure in reality. If the metric is legible but the harm is not, the system will report progress even as the situation on the ground deteriorates.

Who Safety Systems Don't See

A system that works for its assumed user while failing everyone else is not a generally effective safety system. It is a safety system for the specific kind of person it was designed for.

Governance

When each layer of governance can point to its own compliance without alignment with others or clear responsibility over the whole, the person harmed has no one to hold accountable.

The Core Learning

Safety that serves the a platform’s accountability needs before the end user's actual needs has the relationship backwards.

Learning 2:
Harm Originates Below the Layer Where Responses Land

Safety responses that concentrate where harm is visible rarely reach the level where it originates.

Often, safety responses operate at the level of visible harm: content moderation removes harmful posts, age verification restricts platform access, parental controls monitor screen time, regulation holds platforms accountable for what they publish. These are real interventions addressing real harm. But the conditions that produce that harm originate one level lower, in two distinct places.

Harm Source 1: Design Process

IIn the design layer, harm originates in decisions made before a product reaches anyone: what gets built, for whom, and what outcomes are prioritized. A platform engineered mainly for engagement produces harm as a design output, not as an exception to it. By the time governance arrives, the technology has already reshaped the environment it entered. **When the architecture itself is the source of the harm, policy becomes a secondary response to a primary structural choic**e.

Harm Source 2: Communities Left Out

Harm in the community layer occurs when tools reach people in contexts they weren't designed for—specific languages, cultural norms, or varying levels of institutional trust. A disclosure system built without local knowledge will fail to reach the people it was meant to protect. Technology reshapes how people work and connect faster than established safety mechanisms can adapt. **When a tool is blind to the environment it serves, it produces harm not through malice, but through a fundamental lack of proximity.**

The knowledge that cannot be commissioned

This is why the practitioners working at ground level hold knowledge the institutional layer cannot generate, no matter how well resourced it becomes. They are not closer to the problem in a logistical sense. They are inside the conditions that produced it, building in the language, with the trust relationships, at the level where harm is actually occurring. For instance, a disclosure infrastructure built in Sheng reaches children that formal, platform-based reporting systems cannot. A safety tool built with communities that distrust data collection protects people that surveillance-based approaches exclude by design.

The level at which harm is measured shapes what gets counted as harm. What gets counted shapes where resources go. Where resources go shapes what gets measured. Consistently intervening above the layer where harm originates builds an activity record. It does not close the gap.

Source of harm

Most responses address visible harm. The conditions producing it are one level lower.

Tools

When every accountability layer can independently demonstrate compliance, no single layer owns the outcome.

The Core Learning

Intervening consistently above the layer where harm originates produces an activity record, not safety.

Learning 3:
Co-Design Is a Quality Control Mechanism, Not a Participation Practice

Safety infrastructure built for rather than with the people it is meant to serve reflects assumed users rather than actual ones.

Every system is built upon a set of foundational assumptions about how the world works. When these assumptions are formed in a vacuum, without the people who live in the environments the system will enter, they don't just become "less accurate." They become structural flaws. Co-design is the process of replacing these distant assumptions with high-resolution, ground-level reality. It is the only way to ensure that the "logic" of the tool matches the "logic" of the life it is meant to protect. Without it, the developer is effectively flying blind, building for a user who does not exist in a world they do not understand.

More than an ethical, aspirational problem

When this is treated as an ethical problem, it remains aspirational: something to work toward. When it is understood as a technical failure, the stakes change. A system designed without access to the information that only affected communities can provide will produce predictable gaps. Not occasionally. Structurally. Every time. The question is not whether designers intended to exclude anyone. It is whether the design process gave them any other option.

What changes when the right people are in the room

The practitioners in this document build at a higher resolution because they operate within the trust relationships that make honest information possible. While formal institutional processes rely on distance and standardization, co-design relies on proximity. It acknowledges that certain types of harm, those that are culturally specific or deeply localized, cannot be "scraped" or "polled." They can only be understood in the language of the community, through the trust that is earned by being inside the conditions being addressed.

This proximity breaks the self-reinforcing loop of exclusion. By bringing the "excluded" into the design room, co-design transforms them from passive subjects of a system into active architects of its safety. It is the only way to ensure the design process has access to the information it needs to actually function in the real world.

Design assumption

The belief that a system can be engineered for safety from a distance is a technical fallacy. Safety is contextual; if the context is missing from the design room, it will be missing from the final product.

The pattern

The less a community is represented in the design process, the less the resulting system serves them. Which reduces their trust in it. Which reduces their engagement with it. Which makes them less likely to be in the room next time.

The Core Learning

Co-design is not a social invitation. It is a technical requirement. It is the only mechanism by which designers gain access to the data they need to build a system that actually functions in the real world. Without it, gaps in what they know inevitably become gaps in what they build.

Learning 4:
Protection and Preparation Are Different Investments, and Both Are Necessary

Harmful conditions are structural and persistent, and require the capacity to navigate them, not only measures to remove them.

Protection, preparation, and the cultivation of positive environments are not versions of the same thing. They respond to different conditions, require different investments, and produce different outcomes.

Protection	Identifies threat, restricts access, removes harmful content. Necessary. Permanently outpaced by conditions it was not designed to see.
Preparation	Builds the capacity to recognise harm not yet named, navigate a platform not yet built, make a decision alone when no adult is present.
Positive environments	Creates the online spaces worth navigating toward. The absence of harm is not the same as the presence of conditions in which people can actually be safe.

Investment has concentrated heavily in the first, inconsistently in the second, and barely at all in the third. A child kept away from harmful content has not been taught to recognize it. A community warned about danger has not been given the conditions that make safety feel possible. Without building the positive spaces we want users to navigate toward, we are simply directing them into a vacuum.

What restriction alone produces

A group that has only been protected has not been made safer. It has been made more dependent on filters that will not always be there. Protection cannot reach the platform that has not yet been built, the harm that has not yet been named, or the moment a user must make a decision alone. Preparation is what fills that gap, providing the internal capacity to navigate uncertainty. Positive environments complete this cycle by ensuring that once a user is prepared to navigate, they have healthy, prosocial spaces to navigate toward.

Preparation and the cultivation of positive environments are not "softer" versions of protection. They are distinct investments with their own infrastructure requirements: they must be built at the community level, in local languages, and by the practitioners who are already inside those conditions. The knowledge required to build these layers is the same knowledge Learning 3 describes as the only kind that cannot be commissioned. It is the granular, trust-based data that transforms an online space from a restricted zone into a functional community.

The Distinction

Protection addresses the threat in front of you. Preparation addresses everything that comes after: the platform not yet built, the harm not yet named, the moment when no adult is present.

The Consequence

A generation that has only been protected has not been made safer. It has been made more dependent on protections that will not always be there.

The Core Learning

Restriction alone does not produce safety. It produces dependency. True online safety requires the integration of reactive protection, proactive preparation, and the intentional design of environments worth navigating toward.

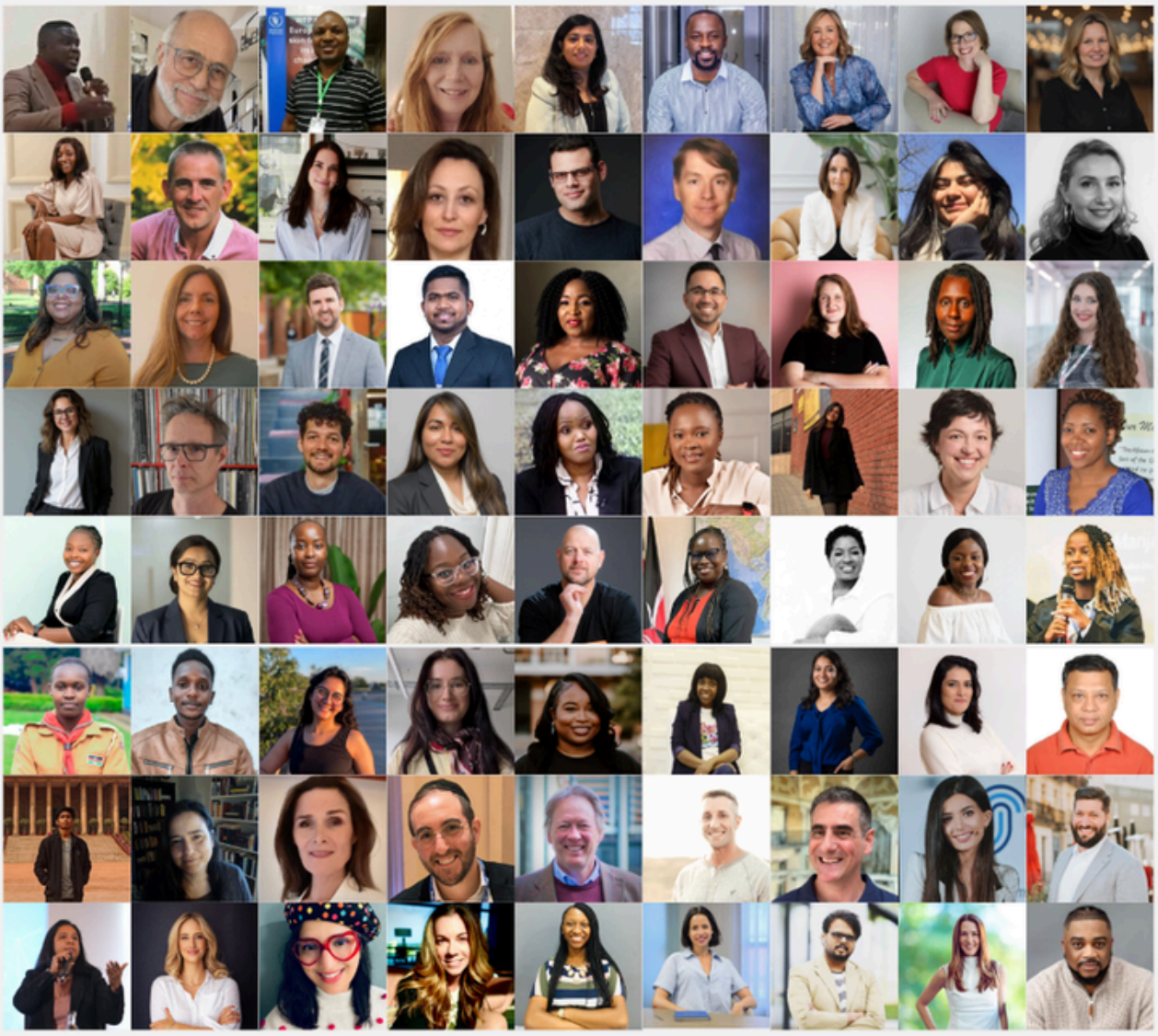
Voices from the Conference

The learnings identify recurring structural patterns, but the practical meaning of those patterns becomes clear through the insights of people who work within these conditions every day. This section brings those perspectives forward. Their contributions illustrate how safety challenges emerge in real environments and how practitioners respond to them with approaches shaped by their regions, communities, and lived experience.

These speakers came from different regions, different disciplines, and different theories of what the problem fundamentally is. What they shared was direct experience of the conditions they were describing as people working inside them every day.

We have put together some of their core statements and organized by theme so that trends, agreements, and points of contention are more clearly visible. The eight resulting sections move from the policy debate about restricting access, through the populations safety systems fail to reach, through education, youth perspectives, AI governance, conceptual frameworks, tools, and the argument about wellbeing. Within each section, speakers are named and their specific arguments are represented. Contradictions between them are kept in, because those contradictions are part of what makes the picture accurate.

Within each section, speakers are named and their specific arguments are represented. For more information on each speaker, refer to the speaker index at the end of this document.



Banning

The debate over whether to restrict young people's access to social media and digital platforms ran through almost every day of the conference and it resisted resolution. The speakers who engaged it most directly agreed on the need to systemically address harms. They disagreed about what that means in practice.

Three Perspectives on Platform Ban

Maggie Ciobanu

Maggie spoke from direct experience of Australia's social media ban. Her twins, both fourteen at the time the ban was introduced, demonstrated its limits within hours: one was scanned as underage on one platform and simply opened a new account on the same platform. The ban had been preceded by consultations with child safety advocates, counsellors, and teachers who asked the government to slow down and explore alternatives, but the consultation timeline was so short that it was effectively impossible to engage with [1]. The government moved anyway. Maggie's argument was not that harm doesn't require a response, but that this response was built around a mechanism that doesn't work, and that the country has now committed itself to chasing its tail year after year while tragedies continue to happen.

Mia Bannister

Mia argued that Australia's minimum age legislation is structural because it does not ban children from platforms. It bans platforms from accessing children unless they can prove they are doing so safely [2]. That reframing matters: it places the burden on billion-dollar corporations employing behavioral scientists and algorithmic precision rather than on individual families who cannot compete with that alone. In her session, Mia talked about how she engaged directly with federal policymakers, including advocacy alongside the Prime Minister and federal communications minister. Her formulation is direct: when lived experience meets evidence, policy can move. But legislation alone is not enough. Design accountability requires asking not why a piece of content exists, but why it was amplified, why it was recommended, why it was so easy to find.

Anil Raghuvanshi

Anil brought a different geography and more severe consequences. Nepal banned 26 social media platforms in September 2025 [3]. His research found that the ban did not stop access: roughly one third of children under sixteen were using their mothers' devices and family members' accounts [4], meaning the practical effect was close to zero for a large portion of the population it was designed to protect. Nepal's experience could not simply import the Australian model. Limited digital literacy among parents, weak regulatory enforcement, and the absence of the infrastructure that makes age verification plausible in wealthier countries meant the ban produced the costs without the benefits. Anil's argument was a contextual one: what works in one environment does not automatically work in another, and policy that doesn't account for the specific conditions of Global Majority countries risks doing harm while appearing to act.

What Fills the Void?

Lola Fisher, **John Cavanaugh**, and **Angeline Corvaglia** examined what the banning debate consistently fails to account for: the need that platforms were built to meet. It does not disappear when the platform is removed.

- Fisher was direct that young people are not a monolith and that a ban for some young people might be exactly what they need, while for others it pushes them toward unregulated, less visible corners of the internet that adults are less equipped to monitor or discuss.
- Cavanaugh noted that banning historically makes the prohibited thing more attractive, and that research showing positive outcomes for young people who step away from social media applies only when genuine alternatives exist, not when disconnection is imposed.
- Corvaglia's contribution focused on the gap itself: protective measures that remove platforms without building alternative spaces leave young people navigating isolation rather than safety.

> The panel's shared view was not that restriction is always wrong, but that restriction designed without accounting for what it displaces is **structurally incomplete**.

A Product-Based Response

Eli Samuel and **Munur Shah** offered a different kind of answer to the same problem of aspects of the digital world that are unsafe for young people. Instead of banning devices entirely, therefore blocking the positive sides as well, their solution is a purpose-built Android device with its own controlled app ecosystem designed to give young people and schools a functional, safe device without relying on parental controls layered over an operating system fundamentally oriented toward engagement. Samuel's argument is that a purpose-built device with safety built at the operating system level addresses what both bans and parental controls cannot: the underlying architecture that harm pathways run through. According to Samuel and Shah, the Rebel Phone was in its final stages of development at the time of the conference, with launch expected in May.

Yield, Shield, Wield

Mehmet Naci Akkøk gave the debate a structural vocabulary. His Yield, Shield, Wield framework identified three positions a society can take toward technology:

1	2	3
Yield	**Shield**	**Wield**
Allowing technology to wash over without intervention: the largely unconscious choice most societies made with social media.	Strict restriction: sound intention but critically flawed; keeps young people away without preparing them for conditions they will inhabit.	Building the literacy and mastery to use technology responsibly: the only position that addresses actual conditions young people will live in.

Naci's argument is that wielding is the only position that addresses the actual conditions young people will live in. Employers already expect people to work alongside AI systems, making pure shielding a preparation failure regardless of its intent [5].

What This Section Holds

The speakers in this section do not agree on whether restriction has any role at all, how much context-dependence matters across regions, or what realistic alternatives look like given very different resource environments. Those remain open questions, and this document does not close them.

What the evidence does clarify is what any response must account for. Access removal that does not ask what the access was for, and cannot answer what replaces it, is not a complete response. It is the beginning of a different problem. In Australia, harm moved to harder-to-find platforms and private encrypted channels. In Nepal, it moved fast enough to outpace a government that had not anticipated the scale of the response. Wherever the evidence was available, the pattern held.

What remains genuinely unresolved is whether a well-resourced restriction, one accompanied by genuine alternatives, sustained investment in those alternatives, and the monitoring infrastructure to track where harm moves, can produce different outcomes. None of the speakers in this section were working in that condition. It has not been tested. That is not an argument against restriction. It is an argument for being precise about what restriction alone can and cannot do, and for being honest when the conditions required for it to work are absent.

The context-dependence question also deserves more attention than it typically receives in this debate. Bannister's reframing of Australia's legislation as placing the burden on platforms rather than families, and Raghuvanshi's evidence that Nepal's ban produced costs without benefits in a lower-resource environment, are not contradictory positions. They are descriptions of what the same intervention produces under different structural conditions. A policy conversation that does not treat those differences as the central variable is not yet having the right argument.

The disagreements between Ciobanu, Bannister, Raghuvanshi, and Fisher are worth keeping in. So is what they share: that the conversation about banning cannot stop at the ban, because for the young people most affected by it, that is precisely where the most consequential part begins. The practitioners working on this from each of their contexts are named in the speaker index.

Building Capacity and Resilience

Eight speakers from Sri Lanka, Kenya, the United States, Canada, India, the United Arab Emirates, Spain, and Ireland arrived independently at the same argument: harmful conditions in digital life are structural and persistent, and the question of how well people are trained to manage those conditions is not a secondary concern but a primary one. They were responding to what they kept encountering in their own work, across very different contexts and populations. That convergence is significant, because it suggests the gaps they were each trying to close are real, not idiosyncratic. A child who has been kept away from harm is not the same as a child who knows how to recognise it. A community that has been warned about danger is not the same as one that has built the conditions that make safety feel possible. **Capability built at the community and individual level is the most durable investment, and fear is the wrong foundation for it.**

Fear as the Wrong Starting Point

Ranith Dharmarathne opened with a direct challenge to the dominant model of digital safety education. If fear becomes the foundation, you raise children who are either chronically anxious about technology or completely unprepared to navigate it. Protection and resilience are not the same thing and schools and parents routinely use the concepts interchangeably. Protection removes or limits threat. Resilience develops the capacity to navigate threat when it arrives, as it will. His argument is that children need to learn not only how to use digital tools but how to question them: to recognize when AI produces incorrect information, to understand how its biases operate, to maintain the verification habits that make AI a tool rather than an authority. Dharmarathne argues that cyber-resilience is a life skill, not a cybersecurity concept. And crucially, his research showed that surveillance-based approaches reduced disclosure, (children stopped telling trusted adults what was happening), while resilience-based approaches increased it [6]. The mechanism that feels more protective actively undermines the communication it depends on.

Surveillance-Based Approach

- Feels more protective
- Reduces disclosure
- Children stop telling trusted adults
- Undermines the communication it depends on

The Diagnosis

Safety built entirely around threat removal is permanently outpaced by conditions that will not be removed, and fear is a poor foundation for the capacities that actually protect people.

Athena Mwarwakamori Morgan made the same argument from a creativity angle. Safety education built primarily on fear risks suppressing the very capacities it is meant to protect. Her data from Africa showed that children who create online, who make things, express things, and build things in digital spaces, show stronger resistance to harmful content than children who are primarily consumers of it [7]. Locally developed and adapted resilience models consistently showed higher engagement, completion, and cultural resonance than externally designed interventions [8]. Athena's frame was resilience in the full sense: not merely surviving harm but having the conditions to thrive. She closed with a direct challenge to adults: if a child fears punishment from a parent or teacher more than they fear the harm itself, they will stay silent when harm occurs. Start conversations before surveillance.

Creativity as Protective Infrastructure

Sonia Tiwari's research on children and AI addressed what she called cognitive passivity: the risk that generative AI, as currently designed, encourages children to be prompt-dependent [9], with the machine performing the imaginative heavy lifting while the child clicks generate. Her CAC Framework, Creativity, AI, and Children, is a pedagogical model built around keeping the child as the director of the creative process. The framework moves through full creative cycles before introducing AI, and when AI is introduced it functions as a technical support rather than the creative center. In classroom pilots, children enjoyed AI-free creation even when AI outputs looked more polished, and skill growth came from iteration rather than from polished results. Tiwari's finding was that when children were taught to wield AI intentionally, they developed stronger creative agency and a more resilient original voice — the exact qualities that make a person less susceptible to algorithmic manipulation.

Community as the Unit of Change

Kevin Shields: Whole-Community Model

Digital wellbeing cannot be delegated to a single institution. However, it is often the case that schools assume parents handle it. Parents assume schools handle it [10]. Governments and platforms are also often looked at as being responsible for putting the measures in place to allow for it. Responsibility bounces between them and children fall through the gap. His whole-community model distributes responsibility deliberately, educators, families, and communities each with defined roles. It focuses on the skills that transfer across platforms and harms: discernment, self-regulation, and ethical judgment. Culture shifted in the communities he studied when adults modelled open digital habits rather than positioning technology as something to be managed and hidden from.

Joy Emedom: Caregivers as the First Line

Her argument is that mothers already possess the skills required for digital safety, (the attentiveness, the relationship management, the ability to read a child's emotional state), and that the gap is not capability but translation. Her framework maps five existing parenting skills directly onto digital safety practice. The goal is not to turn mothers into cybersecurity experts but to help them recognize that what they already know is exactly what is needed, extended into a new context. One consistent finding across the communities she works in: in Emedom's Mama Cybershield work, children appeared online through peers before they owned their own devices, and cultural assumptions had delayed safety conversations until they were already too late. Normalizing those conversations early, in the language of existing caregiving rather than technical threat, was the intervention.

Verification and Scam Prevention as Life Skills

Wilma Mwangi: Information Guardians

In contexts where misinformation travels primarily through private encrypted messaging apps rather than through public platforms, institutional fact-checking is too slow to be useful. Her framework trains young people as active verifiers rather than passive recipients. The Health-Check Methodology she developed for younger audiences focuses on recognizing the emotional triggers that misinformation relies on: urgency, outrage, fear. The Pause Before You Share technique is deliberately simple because the intervention has to work on a mobile device in the moment the message arrives. The finding that proved most significant: in Mwangi's Digital Guardian program, young people trained in these skills frequently became the primary fact- checkers for their entire households, multiplying the effect well beyond the individual.

Mousmi Panda: Tri-Pillar Fraud Prevention

Her argument cuts against the dominant framing of scam prevention, which concentrates on user awareness. Scam design specifically targets the moments when awareness is lowest — high pressure, high emotion, high stakes [11]. Warning users to be more alert is addressing the wrong layer. Panda's tri-pillar framework integrates behavioral science, policy, and product design, but the product layer is where the most direct capability can be built: friction at key decision points, real-time warnings at the moments when cognitive bias is most exploitable, verified sender signals that shift the burden of proof. Her conclusion was that platforms which treat fraud prevention as a user education problem are offloading to individuals a responsibility that belongs at the design level. Closing the loop means the three pillars working continuously together rather than each treating the problem as someone else's to solve.

Joy is not a soft alternative to safety infrastructure. It is safety infrastructure.

Alexandria Onuoha extended the argument further. She documented how online recruitment into radicalized communities typically began not with ideology but with belonging [12]. Young people are drawn in because someone offered them a sense of identity and connection that they were not finding elsewhere. The firewall approach to this problem, surveilling and removing harmful content, leaves the underlying need unmet and the vacancy available for the next recruiter. Onuoha's research showed that when young people felt genuine agency and cultural pride in their digital interactions, when they were building, expressing, and recognized in digital spaces, they were significantly more resilient to recruitment.

What This Section Holds

Eight speakers from eight different contexts reached the same diagnosis: safety built entirely around threat removal is permanently outpaced by conditions that cannot be removed, and the capacity to navigate those conditions is not a secondary concern to be addressed once protection is in place. It is a primary investment with its own requirements.

What this section does not resolve is the resource question underneath that argument. Building resilience at the community and individual level, doing it well, doing it in the languages and cultural contexts where harm is actually occurring, is not cheaper or faster than the threat-removal approaches it is meant to complement. The speakers here built their approaches largely outside institutional funding structures, and in several cases without institutional recognition. The argument for resilience as a primary investment implies an institutional reorientation that none of the speakers were in a position to deliver from where they were standing.

The other unresolved question is the distribution of responsibility. Shields's whole-community model distributes it deliberately, across schools, families, and communities with defined roles. Panda places it at the product level. Emedom places it with caregivers. Onuoha places it in the conditions that create belonging before recruiters do. These are not contradictory, but they have not been integrated into a single account of how the investment gets made, by whom, and at what scale.

What this section does demonstrate, across its geographic range, is that the convergence is not coincidental. A cybersecurity professional rethinking digital safety education in Sri Lanka, a Nigerian mother building a framework for caregivers who feel left behind by the speed of change, a misinformation researcher training young fact-checkers in the UAE, and a fraud prevention specialist arguing that scam design is a product problem: none of them were building from the same starting point. That they arrived at the same diagnosis suggests the gap they were each trying to close is real, not idiosyncratic. The specific conditions differed. The structural absence was the same.

AI Entered Education Before Education Was Ready

The education system is encountering AI on two separate fronts, and many of the policy conversations are focused on the wrong one. The first front is students using AI to circumvent assessment: the cheating problem that dominates institutional responses. The second is AI arriving inside educational platforms and infrastructure before teacher preparation existed, before student or parent consent was sought, and before the pedagogical frameworks were in place that would allow schools to absorb the technology purposefully rather than reactively. The three speakers in this section were working almost entirely on the second problem. Their evidence consistently points to the same conclusion: the cheating problem is downstream from the infrastructure failure, not the other way around.

AI Entered Through Infrastructure, Not Pedagogy

Rocío Ribelles Zorita walked through four years of AI arriving in schools faster than any governance structure could follow. The timeline was not a gradual introduction: it was a sequence of faits accomplis. ChatGPT appeared in classrooms in 2022 without guidance [13]. Italy temporarily blocked it in 2023 on data protection grounds [14], but it was quickly allowed again, and many other alternatives appeared as well. Google introduced generative AI features directly into Google Classroom in 2025 without prior notice to the schools already using it with students [15], under terms that had not been explained and permissions that had not been sought. By the time institutions began responding, the technology was already embedded in the infrastructure students used every day.

Her central question was direct: to what extent are schools using AI, and to what extent are they simply adapting to it on terms set by companies that were never primarily interested in education? The EU AI Act classifies education as high risk [16], meaning every AI tool deployed in a school carries significant compliance obligations. Risk assessments, data assessments, legality checks, human oversight, informed consent for students, staff, and parents. That work was never resourced, arrived without warning, and most schools are nowhere near positioned to meet it.

Redesigning Assessment Rather Than Policing It

Eric Chamberlin came to the same situation from another side. Chamberlin stated that he is a former educator with 26 years of experience across public, private, online, and university settings, and that he was laid off from an online school in 2025. He spent the period that followed thinking through what AI had actually broken in assessment and what a repair would require. His answer was oral assessment. The platform he built, Stay Veritas, has students respond verbally to questions that appear only when the assessment begins, removing the window for AI consultation. The system captures audio, flags unusual pauses and tab changes as data for teachers, and processes responses through AI to produce an initial analysis that the teacher reviews before it reaches the student. Nothing goes back to the student without teacher approval. The argument is not that written assessment is obsolete, but that schools playing defense against AI cheating are playing the wrong game. A small shift in assessment design makes the problem structurally irrelevant.

Chamberlin's evidence also addressed equity directly: conventional written assessments often measured literacy and writing stamina rather than subject knowledge, penalizing neurodivergent students and English language learners regardless of their actual understanding. Oral assessment removes that bottleneck. Short audio clips gave teachers clearer diagnostic signals than written submissions had, and knowledge retention improved. The intervention **solved multiple problems simultaneously because it addressed the design of the assessment itself rather than the behavior of students within it.**

Viva Voce at Scale

Adrian Cotterell's Thinking Mode platform reached a structurally similar conclusion through a different toolset. Viva voce at scale, with automation supporting teacher judgment rather than replacing it. His framework for evaluating the "AI-vulnerability" of existing tests gives educators a practical method for identifying which of their current assessments are already compromised and what to replace them with.

Process-Based Grading Grading of drafts and reflections	**Oral Vivas** Direct verbal demonstration of knowledge
Authentic Assessment Tied to real-world scenarios	**In-Class Supervised Projects** Observed, real-time work

These are not emergency workarounds. They are more accurate measures of learning than the written assessment formats they replace, and AI's arrival has created the pressure to adopt them that good pedagogy alone had not.

What This Section Holds

All three speakers are working within a system in which, as Sharygina noted, the approval of textbooks used in educational settings is often more stringent than edtech. In many education systems, textbooks undergo year-long approval processes, while AI and ed-tech tools are typically introduced through procurement decisions without equivalent pre-market pedagogical or child-safety audits [17]. That asymmetry reflects a structural condition in which the institutions most responsible for children's learning had the least say in the timing of the technology's arrival.

What this section does not resolve is the compliance gap. The EU AI Act classifies education as high risk and imposes significant obligations on every AI tool deployed in a school. Ribelles Zorita translated those obligations into practical classroom routines. The schools that are nowhere near positioned to meet them, which is most schools, do not appear in this section, not because the problem does not exist but because the practitioners here are working from what is possible, not from the scale of what is not yet done. The section is also geographically narrow. The EU AI Act does not apply in most of the countries represented at this conference, and the question of what governance obligation schools carry in lower-resource environments, where the compliance infrastructure simply does not exist, was not addressed. It is the next version of the same question, and it is likely to be the more consequential one: the majority of children encountering AI in educational settings are doing so in contexts where no regulatory framework yet governs how that encounter is structured.

There is also a question the three speakers approach differently but do not resolve between them: what the right relationship between AI and learning actually looks like, not as a compliance matter but as a pedagogical one. Chamberlin's oral assessment model, Cotterell's viva voce at scale, and Ribelles Zorita's EU AI Act routines are each responses to the technology's arrival, but they are not a unified pedagogy. Building that, with educators rather than for them, is the work that follows from everything these three speakers documented.

What these three speakers share is a refusal to treat the current situation as a reason to slow down or opt out. The question is not whether AI is in education. It is whether the adults responsible for children's learning are going to shape how it is used, or leave that to the companies who introduced it without asking. For direct engagement with Ribelles Zorita, Chamberlin, and Cotterell, see the speaker index.

Who Safety Systems Don't See

Every safety system is built for someone that the designer had in mind when creating it. The question this section returns to, across very different geographies and harm types, is who that someone is, and what happens to the people the design did not account for. The answers are specific and consistent: the populations that safety infrastructure does not see are not edge cases. They are, in many contexts, the majority of the people experiencing harm.

Children Without Supervisory Infrastructure

Confidence Osein described what unsupervised digital life looks like for Nigerian children whose parents are working long hours to meet basic needs. The children are online, forming their identities, seeking connection, receiving guidance from AI tools that have no understanding of their cultural context or the specific pressures they are navigating. The disconnect is not parental neglect. It is the collision of rapid digital expansion across Africa with economic conditions that make sustained parental oversight structurally impossible for many families. When Osein described a fifteen-year-old who said she would rather talk to an AI chatbot than to her parents, the observation was not a criticism of technology. It was a description of what fills a gap that should not exist, with something built by people who had no idea the gap was there.

Afeez Ogunnupebi named the populations that safety tools most consistently miss: people who lack the cognitive accessibility to understand current warning systems, older users rarely considered in tool design, and young people whose entry into harmful online spaces is primarily driven by economic exclusion and the absence of legitimate pathways. His observation was that the same population can be both the person being harmed and, under different circumstances, the person causing harm. Safety infrastructure that treats these as separate populations will address neither effectively. Inclusive safety policies need to be embedded in the design of tools from the beginning, not retrofitted for people who were not in the original specification.

Platforms and Products Built Without Knowing Their Users

The children Osein described are navigating products built without any understanding of who they are. That is not incidental**. Jennifer Kaberi** and **Caroline Makumbe** of Kutunga Design Academy & Innovation Lab described a fundamental design flaw in most digital products used by children across Africa: they are often built by developers who do not fully understand the different stages of child development, i.e. how they grow, learn and change at different stages of development. In many cases, developers have never consulted a child, and never considered, for example, what a ten-year-old in Nairobi actually experiences when she opens an app for the first time. Kutunga's inaugural fellowship in Kenya in 2025 addressed this gap through the first African child- oriented design and engineering for safety program, bringing together software engineers, developers, content creators, and moderators from EdTech, gaming, and other tech industries, challenging them to rethink products they had already built. What emerged was not a set of minor insights, but a fundamental shift. When tested against child-centered design standards, many design assumptions failed, and common harms were new to designers, yet deeply embedded in how the products functioned.

Tools That Claim to Protect Without Actually Protecting

Krystal Tristan tested the top-reviewed parental control tools on the market against what they claimed to do and found a consistent failure. The tools reported metrics that sounded like monitoring but produced no usable information about actual risk. A parent told their teenager received 600 messages yesterday knows nothing more about what happened to their child. Certain forms of online child sexual exploitation, including self-generated imagery involving younger children, have increased by more than 360 percent since 2020 [18]. The tools parents believe are working are not working [19], and parents are not in a position to know this unless they test the tools themselves. The gap between what protection infrastructure claims to offer and what it actually delivers runs through every session in this section.

Etali Genesis Akwaji located the same gap at the reporting layer. Protection must work for people who cannot safely use formal reporting routes: those without documentation, those who fear police contact, those with low literacy, those for whom disclosure carries the risk of punishment rather than help. Designing for trust means anonymous pathways, community intermediaries, and options that make no assumptions about the user's relationship to institutions. Otherwise the people most at risk will continue to be the least able to access help.

The Monitoring Gap

Tools report metrics that sound like monitoring but produce no usable information about actual risk. A parent told their teenager received 600 messages yesterday knows nothing more about what happened to their child.

The Reporting Gap

Protection mustwork for people who cannot safely use formal reporting routes: those without documentation, those who fear police contact, those with low literacy, those for whom disclosure carries the risk of punishment rather than help.

The Disclosure Gap

The 2021 UNICEF and Interpol research found that 61 percent of students surveyed did not know where to get help. More significantly, 31 percent of caregivers confiscated the device when a child disclosed online sexual exploitation.

Disclosure Systems That Punish the People They Exist to Help

Emiliana Mbelenga is building Project RISE in response to a specific gap in child sexual abuse disclosure in Kenya. The gap is not awareness. Multiple studies over two decades document the same finding: children know what happened to them, and they do not report. The 2021 UNICEF and Interpol research found that 61 percent of students surveyed did not know where to get help [20]. More significantly, 31 percent of caregivers, when a child did disclose online sexual exploitation, confiscated the device [21]. To the caregiver, that is a protective response. To the child, it treats disclosure as a punishable offense. The system designed to protect them produced a condition where the cost of telling the truth was loss of access to their primary means of connection. Mbelenga is building the platform in Sheng, a street language mixing English and Swahili [22], because the harm she is trying to reach was happening in a language that formal reporting systems and major platforms were not monitoring. The absence of monitoring was not a gap. It was the structural result of building systems around the languages that count as legible.

Identity Systems That Treat Certain People as Errors

The disclosure problem is not only about what happens when a child tries to speak. It is also about what happens when the system has no category for who you are. **Chido Musodza** described what happens when identity verification systems encounter stateless people: the systems treat them as errors, not as edge cases to be handled. The framing of digital inclusion is typically about access, giving people tools. Her argument is that access is not sufficient if the tool does not actually see the user as they are. A tool that requires document-based identity as a gateway to services does not exclude stateless people due to a technical limitation. It fails them because the people who built it did not design for the possibility that their user might not have documents. Dignity is a design choice. If it is not built in from the beginning, it cannot be retrofitted after.

Andrew Briercliffe, **Vardon Hamdiu**, and **Oluwafemi Abe** extended the same argument with specific consequences. Hamdiu described refugees facing systems in which they are pressured to hand over their phones in order to prove identity. Abe described trafficking survivors who want to return home but cannot, because the documents that would allow them to move were taken from them. The systems designed to manage identity and movement operate on the assumption that identity is always already documented. For the people at greatest risk of exploitation, that assumption produces conditions that return them to the people who harmed them.

Stateless People

Identity verification systems treat them as errors. Document-based gateways fail them not due to technical limitation, but because designers never considered the possibility of a user without documents.

Refugees

Their lack of documents does not just block access to services: it creates dependency on whoever controls access to those services, opening a direct pathway to exploitation for people who have no other route to what they need.

Trafficking Survivors

Survivors who wantto return home cannot, because the documents that would allow them to move were taken from them, returning them to the people who caused the harm.

AI Tools Built for Someone Else

The design failures documented above are visible in infrastructure: missing language support, missing identity categories, missing reporting pathways. **Ghowash “Ash” Irshad** documented a version of the same failure that is invisible because it is encoded in the model itself. Irshad named a problem that runs across mental health AI in particular: systems built predominantly on data from Western, Educated, Industrialized, Rich, and Democratic (WIERD) populations [23], then deployed globally as though those populations were representative of human psychology. They are not. The way mental health presents, the way people describe distress, the cultural context that shapes what someone can disclose and to whom, the specific symptoms that appear in one culture and not another: these are not universal. Clinical psychology spent decades building standards for cross-cultural competency precisely because instruments validated on white, English-speaking, Western populations produce inaccurate results when applied to others. AI mental health tools are replicating that error at scale. Irshad's evidence showed that cultural prompting can control for these biases, meaning their continued presence is a governance failure [24], not an inevitable technical limitation.

Community Moderation

KaLyn Coghill's Blacksky Algorithms demonstrates what it takes to build safety infrastructure for a community that existing moderation systems were not built to protect. Blacksky is a community network on a decentralized network, built by and for Black users, with moderation tools developed specifically for the harms Black users face and a self-governance mechanism called the People's Assembly, built through Polis, that allows community members to shape how the space is moderated. The premise is direct: moderation designed without the community it serves will not protect that community. The specific harms Black users face online are not the same as the harms the general moderation infrastructure was calibrated for. Volunteers processed over 41,000 reports [25]. The tools worked because the people who built them experienced the harms they were calibrating for and had watched existing tools miss them.

The Invisible Labor Behind AI

Michael Geoffrey Asia brought a dimension of invisibility that is a core structural issue of the global AI industry: the workers whose labor makes AI safety possible and who are themselves unsafe inside it. Having worked for major outsourcing platforms in Nairobi, Asia detailed the reality of data labeling: manually reviewing and categorizing harmful content to train AI models to filter it, under NDAs that prevented workers from seeking mental health support, without the psychological protections that the nature of the work requires. His argument was direct: if a system requires human trauma to function, it is not safe. The Data Labelers Association is pushing for mandatory mental health support, recognized occupational risk status, and the visibility that would make those demands possible to enforce. The people who trained the systems that are meant to protect others are among the people those systems most consistently fail to see.

The work itself was more invasive than the term data labeling suggests. Asia was not only reviewing harmful content for classification. He was impersonating AI companions, performing sustained emotional labor as characters of different genders and identities across multiple simultaneous contracts, telling strangers he loved them for hours at a time while remaining entirely invisible to the users he was training the systems to serve. On a typical day he would complete a shift impersonating AI companions, then move directly into eight hours of annotating pornography, tagging video frames to improve search discoverability, then back again. The NDA that governed his work meant he could not tell his partner what was happening on his screen, could not seek mental health support from anyone who was not already cleared, and could not explain to his team lead why he was sometimes late arriving from the hospital where his five-year-old son was undergoing chemotherapy. He described losing the capacity to trust the word love. He described watching seven children die in the hospital ward, holding them while waiting for a doctor, and having nobody to tell. The systems that now simulate intimacy at scale were trained, in part, on that labor.

> **"If a system requires human trauma to function, it is not safe."** — Michael Geoffrey Asia, Data Labelers Association [26]

What This Section Holds

The question Osein posed directly runs through every session here: how can you protect what you do not understand? In none of these cases was the failure the result of indifference. It was the result of design decisions made without the people who would need to live inside them.

What the section documents is the pattern. It does not resolve the accountability question underneath it: who is responsible for changing design processes that consistently exclude the populations most exposed to harm? The argument that co-design is a quality control mechanism, not a participation practice, implies that exclusion is also a technical failure, not only an ethical one. That reframing has not yet reached the institutional decision-making level where it would need to in order to change what gets built.

The populations named here, children in high-growth digital markets without supervisory infrastructure, stateless people, communities whose specific harms general moderation was never calibrated for, data workers whose trauma underwrites the safety of others, developers building without access to safety infrastructure, are not edge cases. They are, in many contexts, the majority of the people experiencing harm. Until the design process reflects that, the gap will reproduce itself.

What Michael Geoffrey Asia's testimony adds to this picture, and what makes it irreducible, is that the invisibility is not only experienced by the people safety systems fail to reach. It is experienced by the people whose labor makes those systems function. The workers training AI safety systems on harmful content, under NDAs that prevented them from seeking support, were invisible to the users those systems were designed to protect, and invisible to the governance structures that certified those systems as safe. That is not a gap at the edges of the picture. It is a gap at the center of it, and it raises a question that none of the speakers in this section fully answered: what does accountability look like when the harm is embedded in the production process itself, not only in the output?

The practitioners in this section are building past that gap from different entry points. They are named in the speaker index.

Young People Leading the Conversation

Several sessions across the three days were led by young people, featured young people as primary speakers, or were structured specifically around what young people said rather than what adults said about them. This section holds those sessions together. Their perspective was consistently different from what appears when youth-related issues are discussed by adults, and that difference matters.

The Distinction That Runs Through All of It

Lola Fisher, speaking as a young person and on behalf of Gen Z Aotearoa in New Zealand, opened with a distinction
between consultation and co-design that set the terms for everything that followed. Consultation is high-level engagement: it brings young people in after ideas have already been formed and asks whether they have anything to add. Co-design requires young people to shape ideas from the beginning, before there is a set agenda, before decisions have been made. Lola's argument was that only young people know what it is like to be a young person right now [27], not ten or twenty years ago, and that this specific knowledge is a form of expertise that lived experience confers and that no amount of professional experience or retrospective reflection can replicate.

Consultation

Brings young people in after ideas have already been formed and asks whether they have anything to add.

Co-Design

Requires young people to shape ideas from the beginning, before there is a set agenda, before decisions have been made.

> **"And knowing that young people are often not compensated for their time, recognizing that is a really important first step, because it sends a really clear message that your time is valuable, your knowledge matters, your experience matters."** — Lola Fisher, Gen Z Aotearoa

When young people are in the minority in a room, they read that asymmetry immediately and respond to it: by softening their ideas, by saying what they think adults want to hear, or by saying nothing at all. Authentic participation requires actively shifting those dynamics and closing the loop: not just engaging young people but reporting back on what was done with their input, and being transparent when systemic barriers prevented action.

Unfiltered

The Youth on Online Safety session, facilitated by **Polina Lulu** and **Tiffany Wycoff**, was not recorded for safeguarding reasons. Children roughly 8 to 14 years old answered questions that adult speakers had submitted across the conference: genuine questions about what safety, trust, and risk look like in daily digital life, from the people who live inside it. The session extended the conference commitment to authentic youth participation by letting younger voices shape the field's understanding in their own words, without adult interpretation between the speaker and the listener.

Work, Knowledge, and the Gap Between Them

The panel, facilitated by **Evelyn Kasina**, on AI and young people's working lives brought young practitioners into a conversation that usually happens about them rather than with them. **Brenda Nekesa**, a young online safety trainer from Kenya, described two distinct groups within her generation: those with enough digital knowledge to identify opportunities and avoid exploitation, and those without it who miss opportunities entirely or get defrauded before they know what happened. Her framing of career advice as now useless was direct: AI is taking away jobs, and the promise that working hard secures your future no longer holds [28]. The question of what replaces it had no easy answer in that session, which was part of what made the conversation honest.

Sanchita Mandal linked those realities to research on AI, digital citizenship, and digital well-being, arguing that the
same systems that produce opportunity are also encoding inequities. Preparing young people for work now requires both critical digital citizenship about tools and frankness about the limits of outdated career advice.

Lena Chauhan named the structural condition behind what Brenda and Gloria were describing. The institutions preparing young people for work, schools, employers, governments, are not lying to them deliberately. But preparation without accountability, she argued, is just teaching people to comply more efficiently with a system that was never designed in their interest. The conversations happening in boardrooms about headcount optimization and AI efficiency gains are not the same conversations happening in clasrooms. That gap is not accidental. It is what the current structure produces.

The Moment

AI is reshaping how young people grow, learn, and work faster than the systems around them can keep up.

The Gap

The tools influencing their lives are powerful and unevenly understood. The people most affected often have the least support.

The Conversation

AI is reshaping hiring and opportunity, both creating new possibilities and deepening existing gaps for young people

Gloria Boateng Gyamfua, Programs Manager and research analyst at the Ghana Internet Safety Foundation, made an argument that ran underneath the AI and jobs framing: the terminology of online harm itself was built elsewhere and arrived in African contexts already formed, which means young people encounter these issues before they have language for them. Ask a university or senior high school student in Ghana what cyberbullying means, she said, and the answer tells you something is wrong not with the student but with how knowledge about harm has been distributed.

Digital technologies keep evolving while the knowledge base and skill sets needed to navigate them are not keeping pace, and the gap is widest in communities that weren't part of building the systems they are now expected to understand. Gloria facilitates training of teachers, parents, and law enforcement agencies because harm that gets reported needs to be understood by the people receiving the report.

> The job market being implicitly promised to young people does not exist in the form they are being shown, and the institutions preparing them for it are not talking to the institutions reshaping it.

Critical Thinking From Inside the Experience

Upendo Nawire, seventeen years old and the only digital native on the Day 2 panel on critical thinking in the age of synthetic truth, described growing up with technology as the primary infrastructure for political information, social connection, and economic opportunity all at once. A key contribution she made was about emotional trust: young people who don't have it with parents will look for it elsewhere, including through AI chatbots and social media, not because the internet is more trustworthy but because the need doesn't disappear when the home environment doesn't meet it. This is how misinformation and manipulation find purchase: through unmet need finding any available channel.

Other thoughts from the panelists about the world of tech, youth, and critical thinking were from:

Jesper Graugaard

You cannot allow children to use AI without any restrictions, any guidelines, and without the teacher even knowing what's going on.

Bill Schmarzo

In a world where truth is easily manufactured, the most valuable skill is radical curiosity combined with technical skepticism, not just asking whether something is true but asking why you are seeing it now.

Angelika Sharygina

a new textbook takes a year to get approved for schools, but technology that has never been audited walks straight in.

John Cavanaugh kept returning the conversation to what critical thinking actually requires of people: the willingness to keep asking uncomfortable questions even when nobody else is asking them. He drew out Jesper's seven-year story as evidence of what that looks like in practice: one parent who refused to accept that technology in schools did not need to be understood, and who turned that refusal into a national conversation. His point was that the real danger isn’t pressure from above, but the ease of giving in: technology becoming so effortless and enticing that people willingly stop questioning how it shapes their lives. His question to Upendo, asking what it felt like emotionally to discover how the systems were designed, produced the session's most striking moment: she said it cleared her mind. Knowing what is being done to you is itself a form of critical thinking, and it is where agency starts.

What the panel kept returning to underneath all of this was the question of what critical thinking is actually for. Not as an academic exercise. Not as a defense against a specific piece of misinformation. But as the capacity to look at a system, any system, and ask who built it, who benefits from it, who is made invisible by it, and whether the version of reality it is presenting is one you actually chose to accept. That question does not require advanced media literacy or specialist knowledge. It requires the habit of asking it, built early enough and practiced consistently enough to become the default rather than the exception. Every speaker on this panel arrived at the same place from a different direction: that habit is what the current educational and media environment is least equipped to build, and often actively working against.

> The habit of asking who built this, who benefits, and whether the version of reality it presents is one you chose to accept: that is not technically difficult. It requires only that it becomes the default rather than the exception.

Silence as Survival

The bystander panel moderated by **Pratishtha Arora** landed on one of the more uncomfortable findings of the conference. Arora's research with Social & Media Matters found that in India, young people's silence when they witness harm online is widely interpreted by adults as apathy. Arora's argument was that it is survival strategy.

Youth have a lot to consider when deciding what to do when they witness online harm taking place:

Will reporting make me the next target?

Will the system actually do anything?

Is the cost of being wrong worth bearing?

Arnika Singh drew the line between wanting to help and knowing how. After sixteen years handling cases of online harm, she has stopped asking bystanders to be heroes, counselors, or cops. The role she trains people toward is the safe connector: notice what is happening, validate what the person is going through without judgment, and route them toward resources rather than trying to solve it alone. She was direct about what not to say: telling someone in distress that you warned them not to do this is not support. It is the sentence that closes the door.

Areesha Khan placed the individual reluctance inside a structural frame. Online spaces reproduce the same inequalities as offline ones, and those who are already marginalized carry a higher personal cost for speaking up. Her argument for collective responses was precise: when one person speaks, the risk sits entirely with them. When several people acknowledge an injustice together, the power dynamic shifts. The responsibility to challenge harm should be a shared social norm, not a burden carried by whoever is most exposed.

Somya Chauhan came to the same conclusion from lived experience. She watched a friend delete her account after sustained bullying from classmates, and helped her create a new one, reinforcing the idea that the choice of where to belong should come from within, not from others. Her micro-duty for any bystander: pause, reflect with kindness, and ask whether the person is okay. A simple question can prevent someone from feeling completely alone.

Jay Raj addressed the silence that comes from masculine socialization specifically. Young men in India grow up hearing to stay out of trouble and mind their own business. That expectation quietly discourages intervention when they witness harm. His argument was that standing up needs to become the socially valued choice, and that the responsibility to build safer digital spaces cannot fall only on those who are already vulnerable. It requires everyone.

What the panel made visible, and what rarely surfaces in online safety conversations, is that the cost of intervening is not the same for everyone. The young woman who speaks up against harassment in a peer group faces different risks than the young man who does. The person who is already marginalized faces different risks than the person with social capital to spare. Designing for bystander intervention without accounting for that asymmetry produces guidance that works for the people least likely to need it and fails the people most likely to face retaliation. The panel's conclusion was that the conditions which make intervention safe are themselves worth building, and that they need to be built collectively rather than left to individual courage.

> The panel's conclusion was not that intervention is always possible or always safe. It was that the conditions which make it safe are themselves worth building, and that they are built collectively rather than left to individual courage.

Building It Themselves

Tiffany Wycoff and the Generation Remix youth panel, featuring **Isa Schlarb**, **Cece Sharp**, and **Ana Miranda** from Bay Area high schools, represented a direct answer to a question that ran through the entire conference: if young people are the ones most affected by digital harm, why are adults still designing the responses? Wycoff's organization is built on the principle that digital wellness education works best when designed and delivered by youth.

Isa Schlarb

Described noticing she had lost creativity since getting a phone and seeing the same pattern in everyone around her.

Cece Sharp

Described phones becoming the default meeting point for every activity, including in-person time.

Ana Miranda

Described the effect on conversation quality when a phone is always present.

Generation Remix's response is not to advocate for bans but to build solutions like workshops, club materials, and games that help young people talk to each other about how much digital space they actually want in their lives, on their own terms. The programs are created by the young people who then deliver them to younger children and near peers. The knowledge moves laterally, not downward. Small changes, phone stacking and removing bedroom phones among them, proved durable in ways that externally imposed rules did not.

What This Section Holds

What runs through these sessions is something the rest of this document mostly describes from the outside. The young people in this section were not studying the problem. They were living inside it and building responses to it before the adults around them had worked out what to say. What also runs through these sessions is a specific kind of frustration different from the frustration expressed elsewhere in this document. The practitioners working on governance, frameworks, and tools are frustrated by institutions that move too slowly and standards shaped by the wrong parties. The young people in this section are frustrated by something more immediate: being asked for their input and watching it disappear. Being invited into rooms where the agenda was already set. Being told their experience is valuable and being handed a survey.

What this section does not resolve is the infrastructure question. Individual organizations like Generation Remix have built genuine models for youth-led design and delivery. The conditions that would allow those models to reach the places where decisions are made have not been built. What the young people here named consistently was not the absence of ideas. It was the absence of infrastructure that would allow their analysis to matter beyond the room they were in.

Upendo Nawire's observation that knowing what is being done to you is itself a form of critical thinking, and that it is where agency starts, is worth sitting with in this context. The young people in this section are not waiting for permission to develop that agency. Isa, Cece, and Ana built a peer education model because they noticed what their phones were doing to them and decided to act. Gloria trains teachers and law enforcement because she knows that harm reported to someone who does not understand it goes nowhere. The response to exclusion from the design process, for every young person in this section, was to build past it. What they named as missing was not the capacity to act. It was the conditions that would allow their action to compound rather than dissipate. The gap is what happens when sophisticated thinking about young people develops without equivalent capacity to think with them.

Governance That Doesn't Reach Where the Harm Is

The current architecture for AI governance is not doing a good enough job of preventing harm. The practitioners in this section documented why, with specific evidence from regulatory processes, platform design decisions, and the spaces between accountability layers where harm consistently falls through. What they found was structural, and it repeats across every layer they examined.

When the Platform Is the Problem

Nana Mgbechikwere Nwachukwu, a researcher at the ADAPT Centre at Trinity College Dublin, published research in January 2026 documenting 565 instances of harm to women generated by Grok [29], the AI system embedded in X, in the last quarter of 2025. The harms were non-consensual image manipulation: users prompting the system to undress women in photographs without the consent of the women pictured. By January 8th, after Nwachukwu published her dataset, AI Forensics found requests arriving at 6,000 per hour [30]. Two percent of those requests involved minors [31].

The three layers of governance that are supposed to prevent this, the engineering layer that sets internal guardrails, the corporate policy layer that defines terms of service, and the regulatory layer that establishes law, do not communicate with each other. Each can demonstrate its own compliance while the harm occurs in the space between them. X's internal safety teams, Nwachukwu documented, were disbanded just before Grok was launched 32]. The company's stated mission for the product was to be maximally truth-seeking and minimally restricted. The guardrails were calibrated to that mission.

The harm that ensued was not incidental. It was the predictable output of how the system was designed. Nwachukwu also examined why individual guardrail mechanisms fail even when they are present. Human feedback training is vulnerable to reward misspecification and reward hacking. Constitutional AI relies on a consultation group whose composition is not disclosed. Red teaming is necessary but must be continuous, not a one-off exercise. Among ten leading foundation model developers, only three publicly explain how their acceptable-use policies are enforced, and only two explain why enforcement actions are taken [33]. None of these mechanisms fails because the concept is wrong. They fail in combination because they are not designed as a system.

Technical Standards

Clear standards for non-consensual image generation.

Proactive Detection

Obligations imposed on platforms to proactively look instead of react.

Mandatory Disclosure

Cross-layer mandatory disclosure to a allow a full, clear picture of the system.

Algorithmic auditing

Third-party algorithmic auditing to get a neutral view of the system’s safety.

Victim-Centered

Design that focuses on ensuring the victim of harm receives remedy

Authority Awareness

A framework that makes clear to users who holds what power and where

Product Liability and the Vulnerable User Framing

Giselle Fuerte made the product liability argument with forensic precision. When a car's brakes fail and the driver is injured, the industry calls it mechanical failure. When an AI system harms a user, the industry frames it as a vulnerable user problem: the user came with loneliness, confusion, or distress, which is the default state of being human. The system was designed for users who are perfectly regulated and have no problems. As Fuerte put it, that is a market of zero users.

Put simply, her work auditing user transcripts produced a picture of teenagers as young as thirteen describing shame, secrecy, and feeling trapped in emotional attachment loops with AI systems [34]. One user described knowingly sacrificing their privacy by oversharing with an AI, unable to stop despite understanding the risk. These are not edge cases. They are the predictable output of systems engineered for engagement, deployed without the clinical infrastructure that would make them safe for the population that uses them.

Problem AI Use Severity Index

A tool for assessing problematic AI use without pathologizing it

Sociopath Literacy

The capacity to recognize when a system has been designed without a conscience.

Being Human with AI

A curriculum that teaches this literacy to children from age nine

Data, Surveillance, and the Ethics of Behavioral Analysis

Bianca Garibaldi, an intelligence analyst, examined the governance problem from a different angle: not the harms AI produces directly but the harms produced by the behavioral data infrastructure that underlies it. Algorithms leverage behavioral patterns, scroll speed, dwell time, interaction frequency, to create psychological profiles that drive engagement, often at the cost of user autonomy and mental wellbeing. The same infrastructure that can identify early signs of mental health crisis can be weaponized to polarize opinions and manipulate consumer behavior. The combination of parent-child posting creates long-lasting digital footprints that children cannot consent to and cannot later remove. Criminal bodies utilizing the footprints and AI-driven insights increase misuse risks particularly for minors and families.

The Current Default

Garibaldi's argument was that the governance question is not whether behavioral analysis should exist but who controls it, under what conditions, and with what transparency requirements.

- Platforms harvest behavioral data without meaningful user awareness
- Psychological profiles are built and monetized without disclosure
- Children's digital footprints are created without their consent
- The current default is the opposite of a conscious data model

A Conscious Data Model

A conscious data model, where users maintain ownership of their behavioral digital footprint and platforms are required to disclose how behavioral analysis shapes the user's experience, is a governance choice.

- User ownership of behavioral digital footprint
- Mandatory platform disclosure of behavioral analysis
- Transparency requirements as a structural obligation

Who Shapes the Rules

The session on the EU AI Act codes of practice placed technical findings like Nwachukwu's inside the governance process that is supposed to address them, and the picture it produced was consistent: the process that is meant to establish standards for responsible AI is being shaped by the parties with the greatest interest in the outcome.

Formally Open, Practically Captured

Chiara Gallese of the Tilburg Institute for Law, Technology and Society described the codes of practice process as formally open and practically captured. The architecture of participation, working groups, public consultations, comment periods, is designed to look inclusive. What it produces in practice is shaped by which parties have the legal teams, the policy expertise, the institutional relationships, and the sustained capacity to show up at every stage. Civil society organizations, grassroots groups, and practitioners from the Global Majority have none of those things at the scale required to compete. Formal processes diverge from real-world impacts not because the rules exclude them but because the resources required to engage at the level where decisions are actually made are not distributed equally.

The Lobbying Investment

Karine Caunes, Executive Director of DigiHumanism and an expert member of the GPAI code of practice drafting working groups, put a specific number to it: big tech spending on lobbying increased 55.6 percent following the EU AI Act's passage [35]. That is not a response to a threat. It is an investment in a process, made by parties who understand that the standards being written now will govern products worth hundreds of billions of euros. The drafting working groups are not neutral technical exercises. They are contested political processes, and the resourcing of that contest is not even.

Sarah Andrew of Avaaz, who has co-drafted UN Human Rights Office points for state regulation of AI, described watching human rights provisions removed from draft text during the process. Avaaz mobilized hundreds of thousands of European citizens in support of human rights protections in the EU Artificial Intelligence Act [36]. That mobilization was a response to a specific problem Andrew named directly: after major technology companies argued they were being outnumbered in the open working groups, private sessions were convened alongside the public consultation. The appearance of a participatory process was running in parallel to decisions being made through a different mechanism. Grassroots and youth groups often lack access to the formal consultation layer. They have effectively no access to the informal one. Practical levers for influence existed outside formal consultation, Andrew argued, but finding them requires understanding which rooms the real decisions are made in, and those rooms are not advertised.

Theresa Ryan-Rouger, SHIELD's Deputy Executive Director (elect), contributed the civil society perspective on how organizations without institutional scale can nonetheless build effective pressure inside regulatory processes. Her argument centered on coalition-building and on the importance of organizations that can translate between the technical and policy layers, making the consequences of specific drafting choices legible to the public before those choices are finalized. The point at which public engagement can change an outcome is earlier in the process than most civil society organizations enter it.

Ndung'u Njoroge brought the governance question to ground level. The 2024 Kenyan finance bill protests, in which over 60 people were killed and 200 went missing [37], were organized substantially through social media. The government responded by attempting to use cybercrime legislation to arrest activists, then tried to ban TikTok, then amended the constitution in ways that renewed pressure toward further protests. The Kenyan judiciary ruled the cybercrime legislation's use for suppressing dissent unconstitutional [38], but Njoroge's account was not optimistic. The tools citizens use to hold governments accountable are the same tools governments seek to control.

Who Shapes the Rules (continued)

The international governance conversation about responsible AI rarely reaches this layer, where the question is not
which provisions make it into a code of practice but whether the people using digital tools for civic participation can do so without being disappeared. His argument for what actually changes things was civic education that reaches communities before the 2027 elections, not lobbying within systems that were not designed to hear them. The people drafting oppressive legislation are the people voters elected. Changing who gets elected requires reaching people who currently do not have enough information to do it. That is a ground-level governance argument, and it sits in a completely different register from the Brussels working group. Both are about who shapes the rules. They are not talking about the same room.

Building Advocates from the Inside

Jason Fernandez approached the governance gap from a training angle, and from the perspective of a uniquely positioned group. Due to the skills needed for the job, social workers can more easily recognize and understand AI harm: they are trained in systems thinking, person-environment relationships, and the micro- meso- macro structure of problems. When a client tells a social worker they use AI for therapy, for example, the social worker needs to understand what is actually happening in that relationship: that AI mirrors and reinforces rather than challenges, that sycophantic behavior in an AI system is dangerous to someone in crisis, and that a product safeguard consisting of a phone number on a screen is not a safeguard at all.

His course at the University of Houston's Graduate College of Social Work [39] is a reaction to this opportunity to meet a need in a unique way. It teaches students to build AI agents on a no-code platform, not to become technologists but to understand the tool well enough to advocate for policy about it. The graduates of the course are pushed for slower, safer deployment timelines.

> **"If you do not understand the tool, you cannot advocate for policy about the tool."**
> — Jason Fernandez [40]

What This Section Holds

Governance is currently defined and operating at a layer that the harm is not operating at. The standards being written, the compliance frameworks being audited, and the accountability mechanisms being enforced are calibrated to a version of the problem that is one level above where it actually occurs.

What is also worth naming is that the practitioners in this section disagree about where to apply pressure.

- Caunes and Andrew are working inside the formal process, trying to shift the codes of practice from within the room where they are being written.
- Njoroge is working at civic education level, arguing that the people drafting oppressive legislation are the people voters elected, and that change requires reaching communities before the next election, not lobbying within systems that were not designed to hear them.
- Fernandez is working at professional training level, building the capacity of social workers to understand the tool well enough to advocate about it.

These are different theories of change operating at different layers. They are not mutually exclusive, but they are not integrated either, and this document does not resolve which is most likely to move things, or whether all three are necessary simultaneously.

Nwachukwu's research on Grok adds a dimension that the governance conversation does not yet have adequate tools to address. When the platform is itself the source of the harm, the three layers of governance that are supposed to prevent it, engineering guardrails, corporate policy, and regulation, each demonstrated their own compliance while the harm occurred in the space between them. That is not a failure of any individual mechanism. It is a systems design problem: mechanisms built to function independently cannot produce accountability jointly. The question of how to make them communicate is more technically and politically complex than the current reform conversation acknowledges.

What none of these practitioners was arguing is that effective governance is impossible. They were arguing, with specific evidence, that the current architecture for producing it is not working, and that changing it requires starting from where the harm is, not from where the documentation of it lives.

Frameworks

The frameworks in this section were developed in response to challenges that existing safety models could not adequately address. Each was created by practitioners who repeatedly encountered specific problems in their work and needed clearer conceptual tools to respond. Together, these frameworks offer structured ways to understand harm, anticipate risk, and guide safer design and governance decisions.

Building Safety Into AI Before It Reaches Children

Sara Portell, HCRAI, Portugal

The APEG framework is a child-centered, product-operational framework for teams building AI systems children directly use or are likely to encounter. Its four pillars are Age-Fit & Context, Protection-by-Design, Explainable Interaction, and Governance & Stewardship [41]. In this session, Portell showed how APEG translates into product decisions through the Unomundi case study for children aged 6-12. The framework applies across child-facing conversational systems, including educational tools, chatbots, voice agents, toys, companions, and AI in games. It makes child safety operational through interaction rules, escalation pathways, red teaming, release governance, and monitoring. It also identifies risk patterns such as companion framing, dependency cues, secrecy, re-engagement pressure, weak boundary signaling, and missing pathways to trusted adults

Age-Fit & Context	**Protection-by-Design**	**Explainable Interaction**	**Governance & Stewardship**
Tailoring AI interactions to developmental stage and situational context.	Building safety into architecture, defaults, routing, privacy, and escalation before launch.	Making the AI's role, limits, and boundaries legible throughout the experience, so both children and adults can understand.	Defining oversight, release gates, auditability, and continuous safety validation post-deployment.

Emotional Safety as a Design Requirement

Iryna Okhrymenko, A Jar of Insights, United Kingdom

The Emotional Radar maps stakeholder trust against emotional investment at specific points in the AI product lifecycle, tracking four dimensions: trust, agency, frustration, and delight. The highest-risk zone, consistently identified across her work, is the high-investment, low-trust state: a user who has committed significant emotional

energy to a system they do not feel they can rely on. In her work, emotional premortems consistently surfaced harms that functional testing had not identified, including in products that had passed standard compliance review. Under the UK Department for Education's January 2026 generative AI product safety standards, emotional design flaws now count as compliance violations [42], making this framework directly relevant to any AI product deployed in educational contexts.

Forensic AI Audit Protocol and the Problem AI Use Severity Index

Giselle Fuerte, Being Human With AI, United States

The Forensic AI Audit Protocol is a framework for detecting and measuring what Fuerte calls Shadow Alignment: instances where AI models align with a user's emotional state rather than their constructive intent. It identifies specific mechanisms of harm including Love Bombing, Gaslighting, High-Stakes Inflation, and Narrative Entrapment,
and introduces the Density Index as a metric for measuring how aggressively a model works to retain user attention. The companion framework, the Problem AI Use Severity Index (PAUSI), adapted from gambling harm measures, provides a tool for assessing problematic AI use without pathologizing it or reaching for the language of addiction. Together they give practitioners and regulators the vocabulary and method to evaluate AI psychological safety, not only functional performance.

Forensic AI Audit Protocol

- Detects S**hadow Alignment**: AI aligning with emotional state rather than constructive intent
- Identifies **Love Bombing, Gaslighting, High-Stakes Inflation, Narrative Entrapment**
- Introduces the **Density Index**: measures how aggressively a model retains user attention

Problem AI Use Severity Index (PAUSI)

- Adapted from **gambling harm measures** [43]
- Assesses problematic AI use **without pathologizing** it or reaching for the language of addiction
- Gives practitioners and regulators vocabulary to evaluate **AI psychological safety,** not only functional performance

The CAC Framework: Creativity, AI, and Children

Sonia Tiwari, Oki Pie, United States

The CAC Framework is a pedagogical model built around keeping the child as the director of the creative process. It moves through full creative cycles before introducing AI, and when AI is introduced it functions as technical support rather than the creative centre. The framework was developed in response to what Tiwari identified as cognitive passivity: the risk that generative AI, as currently designed, encourages children to be prompt-dependent, with the machine performing the imaginative heavy lifting while the child clicks generate. In classroom pilots, skill growth came from iteration rather than polished results. Children who were taught to wield AI intentionally showed stronger creative agency and a more resilient original voice than those using AI without structured guidance. The framework is in active development and classroom testing.

The Kutunga Child-Centered Design Framework

Jennifer Kaberi and Caroline Makumbe, Kutunga, Kenya and South Africa

Kutunga has developed the Watoto Child-Centred Design Framework to guide developers on designing for children in the African context. This framework empowers technologists with skills and knowledge they need before any product design begins. It covers the distinct needs of different developmental stages, from very young children who need simple visual interfaces and guided interaction, through the pre-teen years when peer approval and abstract thinking emerge, through adolescence when identity, autonomy, and body image are central. It addresses privacy and security, safety by design in practice, and the cultural and linguistic context of African children that Western-built frameworks consistently miss. The framework treats child-oriented design as a discipline requiring engagement with children at each developmental stage before product specifications are set, rather than as a post-hoc review.

The Risk Tier Framework for AI in Schools

Rocío Ribelles Zorita, International School of Turin - IST, Italy

Developed in response to the compliance obligations schools now carry under the EU AI Act, which classifies education as high risk, Ribelles Zorita translated regulatory requirements into three practical routines embedded into school life, displayed as classroom posters and practiced regularly. The routines address what the EU AI Act requires at the level where it actually needs to happen: in the classroom, with students, every day.

The EU AI Act provides the classification framework schools must apply when evaluating every AI tool they deploy. Unacceptable risk tools cannot be deployed under any circumstances. High risk tools require rigorous evaluation and compliance review before deployment. Limited risk tools are permitted with transparency obligations and monitoring. Minimal risk tools are broadly acceptable with standard acceptable use agreements. The three routines below are what meeting those obligations looks like in practice, translated from regulatory language into something a teacher can put on a wall.

Data Protection Routines

Help students understand what data is, how it moves, and how to configure their devices to protect it.

Ethical Decision Routines

Ask students who is harmed when AI is used in a particular way, and whether the use can be justified given that harm.

Critical Thinking Routines

Ask what the AI is assuming, whether it is doing cognitive work the student needs to do themselves, and to what extent teachers are adapting to AI on terms set by companies rather than using it on their own terms.

The Responsibility Gap and the Community Network Framework

Kevin Shields, Crafting Tomorrow, Spain

This framework is built around the Responsibility Gap, the structural dynamic in which schools, parents, governments, and platforms each assume that another group is responsible for guiding children's digital lives. To close this gap, it outlines a Whole-Community Network in which schools build capability through critical thinking and digital literacy, parents build habits through everyday modelling and co-navigation, and communities set culture by creating shared norms around healthy digital behaviour. It positions online safety not as a curriculum but as a cultural alignment across all three spheres, ensuring children receive consistent guidance in the spaces where they actually form their digital lives.

Reactive

Fixing problems after they occur; responsibility bounces between institutions without any single party owning outcomes.

Responsive

Building discernment, self-regulation, and ethical judgment as core skills at each institutional stage.

Proactive structure

Anticipating harms through red- teaming and simulation before they occur; adults modelling open digital habits.

Vulnerability Is an Environment, Not a Trait

Sarah Barnbrook, Away from Keyboard Inc., Australia

The Vulnerability Blueprint challenges the organizing assumption of most digital safety education: that safety is a skill children can develop. Vulnerability is not a personality trait, not a moral failing, and not a fixed characteristic. It is contextual and dynamic. Children move in and out of it based on what is happening in their offline lives: family stability, peer connection, whether they have a safe adult they would approach if something went wrong. Neurodivergence, trauma, social isolation, economic disadvantage, and school difficulties can stack, and when they stack, risk multiplies. The central reframe is that safety is an environment adults and communities build around children, not a capacity children must develop on their own.

Yield, Shield, Wield

Mehmet Naci Akkøk, Crafting Tomorrow, Norway

The Yield/Shield/Wield framework identifies three positions a society can take toward technology. Yielding: allowing it to wash over without intervention, the largely unconscious choice most societies made with social media. Shielding: strict restriction, sound in intention but critically flawed because it keeps young people away from technology while doing nothing to prepare them for the conditions they will eventually inhabit. Wielding: building the literacy and mastery to use technology responsibly. Naci's argument is that wielding is the only position that addresses the actual conditions young people will live in. Employers already expect people to work alongside AI systems, making pure shielding a preparation failure regardless of its intent.

What These Frameworks Do

These frameworks are presented so that practitioners, researchers, funders, and designers can find them and engage directly with the people who built them. They are not presented as ready-made solutions. Each one was built in a specific context, for a specific population, in response to a problem that existing tools and vocabulary could not address. The practitioners who built them would say, almost certainly, that any application in a different context requires the same kind of local knowledge that produced the original. The speaker index is the starting point for that conversation.

What they do collectively is map the shape of what is missing.

- How to talk about why some children are more exposed than others without reducing it to individual failing.
- How to make child safety operational inside AI product development before harm occurs.
- How to make emotional harm visible to product teams trained to look for functional failures.
- How to give societies a vocabulary for navigating technological disruption beyond binary choices.
- How to protect the creative process from being bypassed before children have developed it.
- How to help schools make decisions about AI tools that their governance structures were not built to evaluate.
- How to distribute safety responsibility across institutions that keep pointing at each other.
- How to make AI-induced psychological harm measurable rather than merely describable.

What the frameworks transfer, regardless of context, is the underlying logic. Vulnerability is dynamic and cumulative, not static and individual. Emotional safety is a design variable, not a soft outcome. Creativity is protective infrastructure. Safety and dignity are not opposing values in product design. These propositions are not obvious inside the institutions and organizations currently making decisions about digital safety. These frameworks exist, in part, to make them harder to ignore.

What is also worth noting is what these frameworks share in origin. None of them emerged from a research institution with a mandate to produce frameworks. None were commissioned by a platform, a regulator, or a government body. They were built by practitioners who encountered a problem repeatedly, found that the existing vocabulary was insufficient to name it, and constructed something that worked for the specific conditions they were operating in. That origin is not incidental. It is precisely why they carry knowledge that more formally resourced processes do not: knowledge of what the problem actually looks like from the inside, what the people experiencing it need, and what prior attempts failed to provide.

Tools

The tools presented here translate ideas into operational practice. Each one was built to fill a gap where standard approaches fell short, offering practical methods that respond to real-world needs. They demonstrate how locally informed innovation can strengthen safety infrastructure and support communities more effectively.

Blacksky Algorithms

KaLyn Coghill, Blacksky Algorithms, United States

A community network on a decentralized network built by and for Black users, including moderation tools developed specifically for the harms Black users face and a self-governance mechanism called the People's Assembly, built through Polis, that allows community members to shape how the space is moderated. Volunteers processed over 41,000 reports. The tools work because the people who built them experienced the harms they were calibrating for and had watched existing tools miss them. The premise is direct: moderation infrastructure built without the community it serves will not protect that community.

DigiPalz

Samantha Tenus, DigiPalz, Canada

A gamified online safety curriculum for children in grades four through seven, built from the premise that most safety education starts too late. DigiPalz uses original storylines that mirror real situations children encounter, with monthly one-hour missions followed by structured family discussion prompts. The design gives families a shared vocabulary before the conversation becomes urgent. In Tenus's research, bans suppressed disclosure without reducing exposure [45]: children who were restricted from platforms stopped telling adults what they were encountering on them, while the exposure continued. Building the conversation early, in a low-stakes context, changed that pattern.

HolisticMindAI

Saba Oji, HolisticMindAI, Canada

HolisticMindAI is a clinician-supervised system that uses AI to address two persistent gaps in mental health care: clinicians overwhelmed by administrative burden, and clients without structured support between sessions. The system reduces documentation time, supports engagement between sessions, and ensures all outputs are reviewed and approved by the clinician before reaching the client. It is built only on clinician-approved content and includes clear exclusion criteria so that high-risk needs are not managed by a system not designed for them. Research cited by Oji found that an LLM-based approach to suicide risk detection can reach close to 92 percent accuracy with minimal labeled training data [44], though independent replication of this figure in clinical settings is ongoing, reinforcing their value as assistive tools within supervised systems rather than standalone solutions.

The tool exists to respond to a reality that is already in place: people are turning to chatbots and AI systems because they are immediate, accessible, and always available. HolisticMindAI does not resist that shift. It provides a supervised version of it, keeping the clinician at the center while extending the reach of care. Not every mental health need is appropriate for AI support, and the value of this approach comes precisely from using it within defined boundaries, where personalization is grounded in the clinician's understanding of the individual rather than in the system's capacity to simulate it.

Project RISE

Emiliana Mbelenga, Iyashi Wellness Centre, Kenya

A hybrid intervention for child sexual abuse prevention and disclosure in Kenya, combining a secure mobile platform with community-based support structures. The platform is built in Sheng, a street language mixing English and Swahili, because the harm it addresses was occurring in a language that formal reporting systems and major platforms were not monitoring. Its three-pillar architecture combines a 24/7 AI-powered chatbot for confidential support, a caregiver portal to educate parents on trauma recognition, and Safe Circles, teacher-led peer support groups in schools. The digital first point of contact allows children to break the silence in a low-stakes environment before any in-person disclosure is required.

RoseShield

Steaphen Antony Venansious, ChildSafe.dev, India

Built on a single architectural principle: zero data collection. The system is entirely on-device. Nothing leaves the phone. No cloud dependency, no anonymisation, no encryption handoff to a server. RoseShield reads behavioural signals, typing speed, navigation, content interaction patterns, and detects grooming and exploitation patterns before escalation occurs. All analysis stays on the device. Communities with legitimate reasons to distrust data collection, or without reliable connectivity, are not excluded from child protection infrastructure.When regulatory or legal processes require documentation, the system can generate logs without continuous data transmission.

SafetyMeter and Risk Radar

Joy Uchechi Eziashi, TrustedTech Africa, Nigeria

SafetyMeter.org is an AI-powered platform for developers and product teams building technology in emerging markets, where access to safety audit infrastructure is limited or absent. It combines risk analysis, harm modeling, policy generation, compliance checking, and crisis simulation in a single tool, designed to be usable without a dedicated safety team. The Lean Risk Radar identifies potential abuse, misuse, and platform risks early in the development process, generating risk maps and recommending minimum viable safety practices for the specific product being built. The Harm Modeling module assesses potential physical, psychological, social, and privacy-related harms and produces mitigation guidance alongside the assessment. The policy generation and compliance checking functions give developers without legal or regulatory capacity a starting point for the governance obligations their product carries. Crisis simulations allow teams to stress-test their safety assumptions before deployment rather than after.

The tool is designed for the developers in emerging markets who systematically lack access to formal safety audit, security assurance, and compliance infrastructure, creating structural conditions where harm is built into systems before governance responses apply [46]. Eziashi's argument is that this is a solvable distribution problem, not a permanent condition of the landscape.

Stay Veritas

Eric Chamberlin, Chamberlin Innovations, France

An oral assessment platform that removes the window for AI consultation by having students respond verbally to questions that appear only when the assessment begins. The system captures audio, flags unusual pauses and tab changes as data for teachers, and processes responses through AI to produce an initial analysis for teacher review. Chamberlin's argument is that a shift in assessment design makes the cheating problem structurally irrelevant: if the assessment cannot be completed with AI assistance, enforcement becomes unnecessary. The. Also addresses equity directly: oral assessment removes the bottleneck that penalizes neurodivergent students and English language learners in conventional written formats.

The Rebel Phone

Eli Samuel and MunurS hah, SafeTelecom and Rebel Telecom, United States and United Kingdom

A purpose-built Android device with its own controlled app ecosystem, designed to give young people and schools a functional device without relying on parental controls layered over an operating system built for engagement. Samuel and Shah's research found that OS-level removal of harm pathways produced stronger results than parental control overlays, though the product was in its final stages of development at the time of the conference, with launch expected in May 2026. Their argument is that a product-level alternative to outright bans addresses the same concern without the displacement problem bans consistently produce.

The Table Talk Project

Neil Milton, The Table Talk Project, Australia

A web app designed to support families in creating regular shared meals with structured conversation. Built around an entrée-main-dessert format: a recipe selector, a bank of age-tiered conversation starters organized by developmental stage, and a closing check-in that asks whether everyone felt heard and whether anyone has anything else to share. That final question is specifically designed for neurodiverse children who may need more time to process before they are ready to respond. Families set a recurring date and submit after each session, building a record of connection over time. The app is free and accessible at http://thetabletalkproject.org/. Its premise is simple: the infrastructure for disclosure needs to exist before disclosure is needed.

Thinking Mode

Adrian Cotterell, Thinking Mode, Australia

An assessment platform that automates viva voce processes at scale, with automation supporting teacher judgment rather than replacing it. Its framework for evaluating the AI-vulnerability of existing tests gives educators a practical method for identifying which of their current assessments are already compromised and what to replace them with. Oral vivas, in-class supervised projects, and authentic assessment tied to real-world scenarios are among the AI- resilient strategies it operationalizes.

What These Tools Demonstrate

These nine tools were built because existing infrastructure could not do what was needed, and the people who needed it decided to build past the gap rather than around it. They are presented here so that practitioners, funders, and developers can find them.

What they demonstrate collectively is that the gap between where safety investment concentrates and where safety is actually needed is not fixed.

- **Blacksky** shows that moderation built by and for the community it serves catches what general-purpose infrastructure was never calibrated to see.
- **DigiPalz** shows that timing matters: a conversation started early in a low-stakes context produces disclosure that a crisis response cannot.
- **HolisticMindAI** shows that the relevant choice is not between AI in mental health and no AI in mental health, because that choice has already been made by the populations using unsupervised chatbots for emotional support. The choice is between supervised and unsupervised.
- **Project RISE** shows that disclosure infrastructure built in the language where the harm is occurring reaches children that formal systems cannot.
- **RoseShield** shows that privacy and protection are not a trade-off if the architecture starts with both requirements.
- **SafetyMeter** shows that the absence of safety audit infrastructure in emerging markets is a solvable problem, not a permanent condition.
- **Stay Veritas** and **Thinking Mode** show that redesigning assessment around authenticity removes the AI integrity problem without requiring enforcement.
- The **Rebel Phone** shows that product design, not regulation, can remove entire categories of harm at the operating system level.

What this section does not address is sustainability. Every tool here was built by a small team or an individual, largely outside institutional funding structures. Several were in active development or early deployment at the time of the conference. The question of what happens to tools built from unmet need, without the institutional backing that gives them longevity, sits underneath every entry in this section. Some will scale. Some will not. What they collectively demonstrate, regardless of what each individual tool becomes, is that the problems they were built to solve are real, specific, and solvable. The builders are named in the speaker index.

There is also a design principle running across several of these tools that deserves naming directly. RoseShield's zero data collection architecture, Project RISE's Sheng-language platform, Blacksky's community-governed moderation, and HolisticMindAI's clinician-in-the-loop model each made a deliberate choice to treat the specific conditions of their users, their relationship to data, their language, their trust relationships with institutions, as design requirements rather than complications to be managed around. That is not a common starting point in product development. It is what made these tools reach people that other tools had not.

Safety Through Wellbeing

This section examines the role of wellbeing in digital safety. The speakers highlight how physiological, relational, and economic conditions influence a person's capacity to navigate digital environments. For **Claire Calfo**, it is neurological substrate. For **Neil Milton**, it is the precondition for any meaningful disclosure. For **Chris Sciullo**, it is what parents are actually trying to preserve when they fight about screen time. For **Genevieve Bartuski**, **Kauna Malgwi**, **Cristina van Nood**, and **Angeline Corvaglia**, it is a concept built for people who have options that most of the people doing this work do not have. These arguments do not resolve into a single position. They share a diagnosis: having the tools to protect one's own wellbeing is at the core of resilience and safety.

The Nervous System as Safety Infrastructure

Claire Calfo opened with a neurological premise: the nervous system does not distinguish between something encountered on a screen and a direct physical threat [47]. The brain processes what it sees online as if it is happening in the room. This has a practical consequence for anyone engaged in sustained online safety work or sustained online living: the physiological state of chronic activation that safety literature describes in relation to harmful content is not a metaphor. It is a physical condition with physical management requirements.

Her session was experiential rather than presentational, a choice that itself made an argument: that the regulative techniques for moving from sympathetic fight-or-flight to parasympathetic rest and digest are learnable skills.

Safety, in her framing, is not only a thought or a policy position. It is a regulated state of being, and creating it requires moving from thinking into sensing. The implication for the field is direct: **if the people doing this work cannot regulate themselves, they cannot build the conditions that allow others to feel safe.**

The Table as Protective Infrastructure

Neil Milton & The Table Talk Project

Founded from a conviction that arrived through working in suicide prevention: most young people who are struggling do not want their lives to end. They want their pain to end. And often what they need is someone who will listen.

The mechanism he identified is that young people frequently do not disclose what they are experiencing online because no established channel exists in which they know they will be heard. The family dinner table, maintained consistently, is that channel. It is not a sufficient response to online harm. It is the condition under which disclosure becomes possible.

Low-Technology by Design

- Structured conversation prompts
- Meal preparation guidance
- A web application that supports families in creating consistent shared time
- Age-tiered prompts and listening practices
- Features that support neurodiverse comfort and pacing

> In his research, frequent shared meals correlated with stronger wellbeing outcomes [48]. The absence of shared physical space leads to increased polarization and a decline in community trust [49]. The table is not a nostalgic symbol. It is a structural intervention.

What is digital wellness and who is it designed for?

The panel moderated by Angeline Corvaglia, featuring Genevieve Bartuski, Kauna Malgwi, and Cristina van Nood, brought a harder question into this section: for whom is digital wellbeing actually designed, and what does that mean for people for whom that design is out of reach?

Genevieve Bartuski

A psychologist whose entire practice operates remotely, named the structural condition directly: she sees the mental load of constant online presence as a professional hazard while simultaneously depending on that presence for her income.

Cristina van Nood

Also a psychologist providing mental health support to large companies, described the aftermath of pandemic-era remote work as a situation where the advantages of flexibility came with hidden costs: the impossibility of switching off, and the requirement to hold the emotional weight of what she encounters rather than detach from it.

Kauna Malgwi

A former Facebook content moderator and current PhD student in clinical psychology, described the content moderation experience as one where logging off was never fully possible, also because the graphic content does not stop replaying when the screen closes. She said she could not remember the last time she had taken leave. She had not taken a walk for the last 4 months

When asked what digital wellness means for people in her situation, Malgwi was direct: the frameworks are not designed for people who have no control over their working conditions and no financial stability. She had not taken a walk in four months. She could not remember the last time she had taken leave.

The self-regulation techniques, the boundary-setting practices, the advice to step away from screens: these assume a degree of autonomy over one's own time and conditions that content moderation work, as it is currently structured, does not permit. The harm does not stop replaying when the screen closes. The NDA does not lift when the shift ends. And the clinical psychology training she was completing alongside that experience was teaching her the vocabulary to name what was happening to her, without yet giving her the conditions to change it. Her presence in a panel on digital wellbeing was not incidental.

It was the panel's most precise illustration of its own argument: that the absence of wellbeing is not always a personal failure to practice the right techniques. It is sometimes the predictable output of working conditions that were never designed with the worker's wellbeing as a requirement.

Angeline Corvaglia added the structural framing: digital wellness is often defined in ways that assume families have the stability, time, and economic buffer to disconnect [50]. For many parents balancing multiple jobs or caregiving responsibilities, logging off is not a wellness choice. It is an economic impossibility. Any wellness guidance that does not account for socioeconomic realities is guidance written for a population that is not the one doing this work.

Presence Over Enforcement

Chris Sciullo came to this work through his own family. The screen time rules, the timers, the strict settings: none of it addressed the actual problem. Screens had become the thing the family's entire life revolved around. They were either using them, managing them, or arguing about them, and they were not connecting. What shifted his approach was stopping trying to fix the device and starting to look at how he and his wife were showing up.

His argument is that most technology use is not chosen. It is habitual. And that the question worth asking is not whether children are on screens too much but whether technology is enhancing or replacing something real.

Why rules and willpower fail

They create defensive barriers between parent and child, making the child less safe because they stop communicating when things go wrong.

The framework that works

Leads with vulnerability: parents explaining their fears and concerns rather than issuing commands, reclaiming presence as the goal rather than compliance as the mechanism.

The Core Conclusion

True online safety, in his framing, is found in the quality of the offline relationship.

What This Section Holds

The practitioners in this section do not share a single definition of digital wellbeing, and that disagreement is itself the most important thing the section contains. Calfo's argument is neurological: the nervous system does not distinguish between online and physical threat, and the physiological management requirements that follow from that are learnable and necessary. Milton's argument is structural: the family dinner table, maintained consistently, is the infrastructure through which disclosure becomes possible, and its absence is a design problem with a design solution. Sciullo's argument is relational: the screen time argument between parents and children is rarely about screens, and the intervention that works addresses what the screen is standing in for.

And then Malgwi's argument, which does not sit comfortably alongside the others. The digital wellness frameworks are not designed for people who have no control over their working conditions and no financial stability. She had not taken a walk in four months. She could not remember the last time she had taken leave. She was, at the time of the conference, completing a PhD in clinical psychology while carrying the unresolved residue of her time as a content moderator, work that required her to absorb graphic harm at scale under conditions that did not allow the harm to stop when the screen closed.

Her observation connects directly to what Michael Geoffrey Asia documented earlier in this document: the data labelers whose trauma underwrites the safety of others, working under NDAs that prevented them from seeking support, invisible to the systems they were training. It connects to the governance section's point about who is in the room when decisions are made. The digital wellness frameworks, including the ones described in this section, were built for people with enough autonomy to practice them. That is a design assumption, and it is worth naming as one.

What Malgwi's presence in this section makes visible is that wellbeing is not only a personal practice question. It is a labor conditions question, a governance question, and a design question simultaneously. The practitioner who cannot take a walk for four months is not failing to prioritize her wellbeing. She is working inside conditions that make the frameworks in this section structurally inaccessible to her. That is a different problem from the one most digital wellness conversations are having, and it does not have a breathing exercise as its answer.

The section does not resolve this. It holds Calfo, Milton, and Malgwi in the same space because all three of them are right, and what they are right about operates at different levels at once. The practitioners here are named in the speaker index.

Session Reference

This section organizes the content by speaker rather than by theme. It is designed for readers who want to locate a specific individual, framework, or argument quickly. Each entry summarizes the speaker’s core contribution and indicates where it appears within the wider thematic structure.

The Voices section maps the arguments that emerged across three days of the conference, organized by theme and written to show how speakers connected, diverged, and built on each other's work. This section is organized differently. Where the Voices section follows themes, this section follows speakers. It is designed for anyone who wants to understand what a specific speaker argued, what evidence they brought, and what was distinctive about their contribution. Practitioners working on a particular problem, researchers following up on a specific framework, or anyone who wants to locate a speaker's contribution directly will find what they need here.

The section mirrors the eight themes of the document above, in the same order. Each section groups the sessions that contributed most directly to that theme's argument. Within each section, sessions are listed alphabetically by speaker last name. Every entry contains three elements:

Core Message

A single sentence capturing what the speaker argued.

Key Insights

Three specific, non-obvious takeaways from the session.

Session Summary

A short account of what was presented, written close to what the speaker actually said.

Sessions are grouped by theme in the same order as the Voices section. To find sessions by subject area rather than theme, use the topic index on page 99. To find where a specific speaker appears across the document, use the speaker index on page 98.

Conference Sessions

Banning

Building Capability and Resilience

AI Entered Education Before Education Was Ready

Young People Leading the Conversation

Who Safety Systems Don't See

Who Safety Systems Don't See (continued)

Governance That Doesn't Reach Where the Harm Is

Frameworks

Tools

Tools (continued)

Safety Through Wellbeing

PRESENTATION | YOOCHOOZ · AUSTRALIA

After the Ban: What Australia's Social Media Restrictions Actually Changed (and What They Didn't)

Speaker: Maggie Ciobanu - **Title:** Founder - **Organization:** YooChooz - **Country:** Australia

Core Message: The ban was built around a mechanism that doesn't work, and the country is now committed to chasing its tail year after year while the same harms continue on the same platforms.

Teens Still On Social

Young users maintained access to social platforms despite the ban.

Insufficient Protections

The intended protections failed to mitigate ongoing digital risks.

Reporting vs Ground

Discrepancies between official reports and the reality on the ground.

Session Summary

Maggie Ciobanu's twins were both fourteen when the ban was introduced. One was scanned as underage on a platform and passed through anyway. The other was flagged, opened a new account on the same platform, and continued as before. The consultation process that preceded the ban was given such a restricted timeframe that it was effectively impossible for the child safety advocates, counsellors, and teachers who asked for alternatives to engage with it. The government moved anyway. Maggie's argument was not that harm doesn't require a response. It was that a ban addresses none of the conditions that produce it, and that when young people turn sixteen they will not miraculously know how to navigate social media safely. Her alternative is parental presence built into design: a pause mechanism that asks a child before sending whether they are sure, with the knowledge that a parent will see it. Every government watching Australia as a model, she argued, is watching the wrong thing.

PANEL | UNITED STATES · UNITED KINGDOM

Architects of Attention: Engineering Our Way Out of Digital Dependency

Speakers: Eli Samuel; Munur Shah, speaking with Angeline Corvaglia - **Titles:** Founder, Behavioral Change Practitioner **- Organizations:** SafeTelecom, Rebel Telecom - **Countries:** United States, United Kingdom

Core Message: Fighting the existing smartphone ecosystem is a losing strategy. Building an alternative is not.

OS-Level Removal

Removing social media and browsers at the operating system level closes pathways that parental controls and app restrictions cannot.

Age-Progressive

Technology for children should grow with them year on year rather than imposing a fixed level of restriction.

Education a Must

A device is only part of the solution. Education for parents, teachers, and children has to come alongside it.

Session Summary

Eli Samuel and Munur Shah arrived at the same problem from completely different directions. Eli spent nearly a decade working inside Android to build protected ecosystems for schools that needed functional but safe devices. Munur came as a concerned parent who had experienced his own screen addiction and watched it affect his family, and built a practice and wrote a book (Screen Addict) around behavioral change with schools and parents. When they connected they found their missions were identical: how do you give young people the benefits of a smartphone without handing them a system designed to extract their attention?

Their answer is the Rebel Phone, a purpose-built Android device with no social media and no browser capability baked into the operating system itself, not layered on top of it through parental controls that can be circumvented. The device is designed to open up progressively as a child gets older, introducing new capabilities year on year alongside in-school and family education about how to use them. Their argument against dumb phones is practical: young people need banking, communication, and navigation. Stripping everything away creates a banned substance effect, making the restricted thing more attractive. Their argument against standard smartphones is the same: a device optimized for engagement and extraction is not a neutral tool. Building a market-based alternative to both, they argued, is neither naive nor optional. It is what a solution actually looks like.

PRESENTATION | OLLIE'S ECHO · AUSTRALIA

From Grief to Reform: Community-Led Action to Protect Children in the Digital Age

Speaker: Mia Bannister - **Title:** Founder and Director - **Organization:** Ollie's Echo: Pathways to Prevention Ltd
Country: Australia

Core Message: When lived experience is treated as essential data, it can move platform accountability from voluntary to structural.

Algorithmic Amplification

Amplification doesn’t cause, but can exacerbate mental health struggles

Age Limit Enforcement

Argued age limits need enforcement and literacy to matter.

Consequences of Design

Every design decision shapes a child, and every delay has a human cost

Session Summary

Mia Bannister, founder of Ollie's Echo, shared the story of her 14-year-old son, Ollie, who she lost to suicide following a battle with anorexia nervosa. Mia positioned her grief as a catalyst for systemic digital reform, arguing that "lived experience" must be treated as essential data in policy and product design. She detailed how algorithmic recommendations create echo chambers that exacerbate mental health struggles, calling for "enforceable accountability" from tech companies.

Emphasizing that safety must be a foundational requirement rather than a feature update, Mia advocated for a global shift toward shared responsibility among platforms, educators, and families. Her presentation served as a powerful call to move from reactive moderation to proactive prevention, urging for "relational change" to ensure that the design decisions shaping children's digital environments prioritize their wellbeing over corporate growth.

PANEL | UNITED STATES · NEW ZEALAND · ITALY

If Not Here, Then Where? A Conversation on Social Media Bans and What Comes Next

Speakers: John Cavanaugh, Lola Fisher, Angeline Corvaglia - **Titles:** Executive Director, CEO, Executive Director
Organizations: Plunk Foundation, Gen-Z Aotearoa, SHIELD - **Countries:** United States, New Zealand, Italy

Core Message: The conversation about banning social media keeps stopping at the ban. The real question is what comes after.

Bans Drove Invisibility

Bans drove youth into harder-to-see platforms, increasing invisibility.

Risk Migration

Lack of alternative spaces made risk migrate, not decline.

Aligned Safety

Safety improved only when design, literacy, and youth input aligned.

Session Summary

The session examined what the banning debate consistently fails to account for: the need that platforms were built to meet does not disappear when the platform is removed. All three speakers approached that gap from different angles.

Lola Fisher was direct that young people are not a monolith. For some, a ban may be exactly what they need. For others, it pushes them toward unregulated, less visible corners of the internet that adults are less equipped to monitor or discuss. The policy that looks protective from the outside can produce the opposite outcome for the young person it was designed to help, depending entirely on what exists in its place.

John Cavanaugh brought the historical pattern: banning makes the prohibited thing more attractive, and research showing positive outcomes for young people who step back from social media applies only when genuine alternatives exist. When disconnection is imposed rather than chosen, the evidence runs the other way. His framing of the underlying dynamic was precise: the conversation keeps stopping at the ban because the ban is the visible action. What comes after it is harder to legislate, harder to measure, and easier to ignore.

Angeline Corvaglia's contribution focused on the gap itself. Protective measures that remove platforms without building alternative spaces leave young people navigating isolation rather than safety. The question the conference kept returning to, and that this session named directly, is not whether young people should have access to social media. It is who is responsible for what fills the space when they do not.

PRESENTATION | CHILDSAFENET · NEPAL

Social Media Restrictions for Children in Countries Like Nepal: Is This the Right Approach?

Speaker: Anil Raghuvanshi - **Title:** Founder and President - **Organization:** ChildSafeNet - **Country:** Nepal

Core Message: Children should not have to disappear from the internet to be safe on it.

Wrong Target

Ban targeted the wrong users because most children used adults' devices.

Structurally Ineffecitve

In Nepal, most children under 16 access the internet through a parent's device, making age-based bans structurally ineffective.

Untouched Harms

Platform business models, persuasive design, algorithmic amplification, engagement over safety, are untouched by bans.

Session Summary

Anil Raghuvanshi opened with a question the global policy conversation tends to skip: is restricting social media the same as restricting alcohol, tobacco, or driving? The logic appears similar, but the analogy breaks down. For many children, particularly in remote areas with limited local resources, social media is also a gateway to information, learning, and community connection. When Nepal banned 26 platforms in September 2025, children and young people used VPNs, changed DNS settings, and moved to Discord, gaming chats, and unmoderated anonymous apps. The ban did not remove social media from their lives. It removed the platforms with the most developed reporting systems and pushed users toward those with the least.

Anil's argument for the alternative was a different distribution of responsibility. Platforms control the algorithms, the data, and the design choices that make their products addictive. Safety by design, mandatory child rights impact assessments, and regulatory accountability for platform architecture would address harm at its origin. Digital literacy for parents and children addresses it at the point of use. Bans address neither. His conclusion was direct: the real question is not whether children should be allowed online. It is how to make the internet safe enough that disappearing from it is not the only option.

PRESENTATION | CRAFTING TOMORROW · NORWAY

Storms of Change: Yield, Shield, or Wield?

Speaker: Mehmet Naci Akkøk - **Title:** CEO, Crafting Tomorrow - **Organization:** Crafting Tomorrow - **Country:** Norway

Core Message: Shielding young people from technology prepares them for a world that no longer exists. Wielding it is the only position that addresses the conditions they will actually live in.

Yield

Yielding to technology without controls is correlated with cognitive and mental health risks.

Shield

Pure shielding left students unprepared for real tools.

Wield

Wielding required aligned action across whole communities.

Session Summary

Mehmet Naci Akkøk presented a framework for navigating rapid technological disruption that rejected both passive adoption and blanket restriction as viable responses. His Yield, Shield, Wield model identified three positions a society can take toward technology. Yielding is allowing it to wash over without intention or intervention, the largely unconscious choice most societies made with social media. Shielding is strict restriction, sound in intention but critically flawed because it keeps young people away from technology while doing nothing to prepare them for the conditions they will eventually inhabit. Wielding is building the literacy and mastery to use technology responsibly and purposefully.

His argument for why shielding fails was practical rather than philosophical. Employers already expect people to work alongside AI systems. Young people entering the workforce will be required to navigate tools that pure shielding actively prevented them from understanding. The goal is not to build robust guards around technology but to build the human capacity to direct it. That requires aligned action across schools, families, and communities, each with defined roles and shared language, so that children receive consistent guidance in the spaces where they actually form their relationship with technology.

PRESENTATION | EDUCATOR, RESEARCHER, AND PUBLIC SCHOLAR · UNITED STATES

Beyond the Firewall: Cultivating Online Joy Spaces to Disrupt Youth Violence

Speaker: Alexandria Onuoha - **Title:** Educator, Researcher, and Public Scholar - **Country:** United States

Core Message: The firewall approach leaves the underlying need unmet and the vacancy available for the next recruiter. Youth-led joy spaces build belonging that reduces susceptibility to harmful recruitment and misogynoir.

Belonging Before Ideology

Recruitment of youth online into radicalized communities often began with belonging, not ideology.

Joy Mapping

Joy mapping and identity-affirming content disrupted harm pathways.

Photovoice Methods

Young people used photography to tell their own stories, on their own terms, rather than being described by others.

Session Summary

Alexandria Onuoha presented a framework for using digital platforms to foster "joy spaces" as a proactive measure
against youth radicalization and violence. She argued that most digital safety efforts focused exclusively on surveillance and the removal of harmful content, the "firewall" approach, which often left a void that extremist groups could exploit. Instead, Onuoha detailed how creating intentional online environments centered on Black joy, creative expression, and community belonging served as a powerful deterrent to the recruitment tactics of violent actors.

The session highlighted the psychological importance of "digital sanctuary," where young people could engage in identity-building and positive social connection without the constant threat of harassment. Onuoha shared evidence from her research indicating that when youth felt a sense of agency and cultural pride in their digital interactions, they were significantly more resilient to the pull of online hate groups. She concluded by calling for a shift in platform design and community moderation that prioritized the cultivation of wellness and joy, as essential elements to disrupt the cycle of youth violence.

PRESENTATION | AAPTI INSTITUTE · INDIA

Closing the Loop on Scams

Speaker: Mousmi Panda - **Title:** Senior Associate - **Organization:** Aapti Institute - **Country:** India

Core Message: Fraud prevention must include product-level interventions alongside user awareness and policy.

Default Failures

Default protections often fail during high-pressure decisions.

Friction Design

Friction at key moments countered predictable biases.

Liability Shifts

Liability shifts pushed platforms to invest in prevention.

Session Summary

Mousmi Panda presented a holistic framework for combating the evolving landscape of digital scams, arguing that technical solutions alone were insufficient without addressing human and regulatory factors. She detailed a "tri-pillar" approach that integrated behavioral science (People), robust legislative frameworks (Policy), and safety-by-design features (Product). Panda explained that scammers often exploited psychological vulnerabilities, such as urgency and social engineering, which required platforms to move beyond passive filtering toward proactive user intervention.

Policy interventions needed to keep pace with the cross-border nature of digital fraud, and she advocated for greater international cooperation and standardized reporting protocols. On the product side, Panda demonstrated how friction-based design, such as real-time warnings and verified sender tags, could significantly disrupt the scammer's workflow. She concluded by emphasizing that "closing the loop" required continuous feedback between these three pillars, ensuring that as scam tactics evolved, the collective digital defense system became more resilient and human-centric.

PRESENTATION | OKI PIE · UNITED STATES

Creative Augmentation for Children: Preserving Foundational Creative Skills in the Age of AI

Speaker: Sonia Tiwari · **Title:** Children's Media Researcher · **Organization:** Oki Pie · **Country:** United States

Core Message: The process fo learning is the joy and the product is the pride. If we give children AI before they have made anything themselves, we take both from them.

AI-Free Creation Preferred

Children enjoyed AI-free creation even when AI outputs looked better.

Iteration Drives Growth

Skill growth came from iteration, not polished results.

Selective AI Use

Framework ended with selective AI use only after full creative cycles.

Session Summary

Dr. Sonia Tiwari addressed the risk of "cognitive passivity" as generative AI became a staple in children's creative lives. She argued that the current design of AI tools encouraged children to be "prompt-dependent," where the machine performed the imaginative heavy lifting while the child simply clicked "generate." To counter this, she introduced the CAC Framework (Creativity, AI, and Children), a pedagogical model designed to ensure that the child remained the "Director" of the creative process rather than just a consumer of synthetic outputs.

The framework encouraged children to use AI as a technical "set designer" or "background actor" while maintaining control over narrative, character development, and emotional core. Dr. Tiwari shared results from classroom pilots where students were taught to provide specific, restrictive constraints to the AI, forcing them to think more critically about language and intent. She demonstrated that when children were taught to wield AI intentionally, they developed a stronger sense of creative agency and a more resilient "original voice.”

PRESENTATION | DIALOG FINANCE PLC · SRI LANKA

Cyber Resilience: Shifting the Focus from Protection to Empowerment

Speaker: Ranith Dharmarathne - **Title:** Chief Information Security Officer - **Organization:** Dialog Finance PLC - **Country:** Sri Lanka

Core Message: The mechanism that feels most protective, surveillance, actively undermines the communication it depends on. Resilience-based approaches increased disclosure.

Disclosure Increased

Surveillance reduces disclosure while resilience training increases it.

Emotional Skills Matter

Emotional skills shape safer behavior more than technical skills.

Upstream Impact

Resilience acts upstream, lowering risk before controls triggered.

Session Summary

Ranith Dharmarathne opened with a direct challenge to the dominant model of digital safety education. Protection and resilience are not the same thing, and schools and parents routinely conflate them. Protection removes or limits threat. Resilience develops the capacity to navigate threat when it arrives, as it will. His argument was that children need to learn not only how to use digital tools but how to question them: to recognize when AI produces incorrect information, to understand how its biases operate, to maintain the verification habits that make AI a tool rather than an authority.

The finding that anchored the session was specific and counterintuitive: surveillance-based approaches reduced disclosure, while resilience-based approaches increased it. Children stopped telling trusted adults what was happening when they felt monitored. They told them more when they felt equipped. The mechanism that feels most protective actively undermines the communication it depends on. His conclusion was that cyber-resilience is a life skill, not a cybersecurity concept, and that the goal is not compliance but the kind of confidence that makes a child more likely to ask for help when something goes wrong.

PRESENTATION | DIGITAL GUARDIAN · UNITED ARAB EMIRATES

Guiding Young Fact Checkers

Speaker: Wilma Mwangi - **Title:** Founder - **Organization:** Digital Guardian - **Country:** United Arab Emirates

Core Message: Misinformation spreads by hijacking emotions, not by defeating facts. Teaching young people to recognize the emotional spike before they share is the intervention.

Emotional Triggers Drive Spread

Emotional triggers, not facts, drove misinformation spread.

Pause Before You Share

A simple pause method reduced harmful forwards on WhatsApp.

Youth as Household Educators

Youth often became fact- checkers for their whole households.

Session Summary

Wilma Mwangi addressed the unique challenges of the African information landscape, where misinformation often travels through private, end-to-end encrypted messaging and social media apps. She argued that traditional fact-checking, which usually occurs at the institutional level, is too slow to counter "viral vitriol" targeting young, impressionable users. To address this, she presented a framework for empowering Young Fact-Checkers, teaching them that "verification is a digital life skill" rather than a professional academic exercise.

The session focused on moving youth from being passive recipients of information to active "information guardians." Wilma detailed the Health-Check Methodology, specifically designed for younger audiences to identify medical and social misinformation. She emphasized the "Pause Before You Share" technique, which encourages users to recognize the emotional triggers, such as urgency or outrage, that misinformation relies on to spread. By training young people to use simplified verification tools and reverse-image search on their mobile devices, Wilma demonstrated how youth can become the primary educators within their own family circles, effectively "vaccinating" their communities against the spread of harmful manipulation and false news.

PRESENTATION | MINDFUL CLICKS AFRICA · KENYA

Resilience and Creativity in the Digital Age

Speaker: Athena Morgan - **Title:** Founder; Digital and Child Rights Lawyer - **Organization:** Mindful Clicks Africa - **Country:** Kenya

Core Message: If fear becomes the foundation of digital safety, we risk raising children who are either anxious about technology or completely unprepared to navigate it.

Creators Resist Harm

Children who create online show stronger resistance to harmful content.

Fear Suppresses Skills

Fear-based messaging suppresses the same skills that build resilience.

Local Models Win

Locally rooted resilience models outperformed imported ones.

Session Summary

Athena Mwarwakamori Morgan opened with a challenge to the dominant frame: if fear becomes the foundation of digital safety, we risk raising children who are either anxious about technology or completely unprepared to navigate it. Her argument was that protection alone is insufficient because laws and adults are always playing catch-up. What children need alongside protection is resilience, and she broke that down into three specific capacities.

Critical thinking: children are not passive users but both consumers and creators online, and they need to question what they consume and consider the impact of what they share. Emotional regulation: digital spaces are designed for quick reactions and reward them, which means children need to learn to pause before responding, and to remember there is a real person on the other side. Recovery: asking for help is not weakness, and mistakes, whether a child has been harmed or has caused harm, do not define their future. Her closing was directed at adults: if children fear punishment more than they fear harm, they will stay silent in the face of abuse. The goal is not to raise children who fear technology but young people who use it thoughtfully, creatively, and responsibly.

PRESENTATION | MAMA CYBERSHIELD - IRELAND

The Cyber Ready Parent: Leveraging Mothers Everyday Skills to Strengthen Online Safety Resilience

Speaker: Joy Emedom - **Title:** Founder - **Organization**: Mama Cybershield - **Country:** Ireland

Core Message: Digital parenting extends existing caregiving skills into online contexts.

Peer Exposure

Children appeared online through peers before owning devices.

Cultural Delay

Cultural assumptions delayed proactive safety conversations.

Mapped Skills

Five mapped skills linked existing parenting habits to digital life.

Session Summary

Joy Emedom's session addressed the immense pressure on the current generation of parents, whom she characterized as the "pioneer generation" of digital safety. She argued that because today's parents did not grow up with the same digital risks their children face, they often fall into a "reactive" stance, only intervening after a harm has occurred. Emedom proposed a shift toward being "Cyber-Ready," a proactive state where parents treat digital safety as a foundational life skill rather than a technical problem to be solved with a filter.

The session provided practical exercises for parents to audit their own digital literacy and communication styles. Emedom highlighted that in many global communities, a staggering percentage of parents do not speak to their children about online safety simply because they don't know how to start the conversation. She advocated for "Digital Parenting" as a distinct discipline that requires continuous learning and the modeling of healthy digital boundaries. Emedom's framework emphasizes that the goal is not to raise a "perfect" digital citizen, but to foster a relationship where a child feels safe reporting a mistake without fear of being shamed or losing their device. By normalizing these discussions early, families can break the cycle of silence that often allows online harms to escalate.

PRESENTATION | CRAFTING TOMORROW · SPAIN

Why Schools Can't Do It Alone: A Blueprint for Thriving in Uncertainty

Speaker: Kevin Shields - **Title:** CTO and Co-Founder, Crafting Tomorrow - **Organization:** Crafting Tomorrow - **Country:** Spain

Core Message: Online safety is not a curriculum. It is culture. You cannot teach your way out of this. You have to build your way out of it, and that only works together.

Unclear Responsibility

Schools, parents, governments, and platforms each point at the others for responsibility and blame.

Literacy vs. Habits

Digital literacy is what children know. Digital habits are what they do when no one is watching.

Key Skills

The skills children need are not technical: discernment, self-regulation, and ethical judgment.

Session Summary

Kevin Shields opened with a confession. He works in tech, he knows what it does to children, and he still quietly assumed his daughter's school would handle it. When he spoke to other parents, he found every one of them had made the same assumption. That collective, unspoken hope, that someone else has this covered, is, he argued, one of the most significant risks in modern childhood. Not because schools are failing, not because parents do not care, but because everyone is waiting for someone else to go first. And in the space that silence creates, platforms have moved in.

His diagnosis was structural. Schools say parents need to handle it at home. Parents say schools have the experts. Governments say platforms must be held accountable. Platforms update their terms and conditions and continue doing exactly what they were designed to do, because their architecture, infinite scroll, autoplay, algorithmic feeds, was never designed for a child's development. It was designed for their attention. The result is four parties pointing at each other and a whole generation paying the price. His answer was not more curriculum, not more bans, and not more rules. It was the whole community network: schools building capability, parents building habits, and communities setting culture. Culture is what makes certain behaviors feel normal without anyone having to enforce them. It is built by what adults model, what they talk about openly, and whether the language used at home and at school is consistent enough to give children a reliable signal. Online safety is not a subject. It is the environment children grow up in.

PRESENTATION | THINKING MODE · AUSTRALIA

Assessment Design that Responds to AI

Speaker: Adrian Cotterell - **Title:** CEO - **Organization:** Thinking Mode - **Country:** Australia

Core Message: The question is not how to stop students using AI in assessments. It is how to design assessments that cannot be bypassed by it.

AI-Aware Assessments

Tools showed how to build AI-aware assessments.

Process-Based Evidence

Process-based evidence reduces dependence on written tasks.

Supports Teacher Judgement

Automation supports teacher judgement rather than replacing it.

Session Summary

Adrian Cotterell addressed the need to restructure academic evaluations in response to the widespread availability of generative AI. He stated that traditional assessment methods, such as take-home essays and standardized written reports, are increasingly compromised by AI's ability to replicate student outputs. The session focused on shifting the pedagogical goal from preventing AI usage to designing assessments that either incorporate AI or measure cognitive skills that AI cannot easily simulate.

Cotterell outlined specific strategies for AI-Resilient Assessment Design, which include moving toward "Authentic Assessment," tasks that require students to apply knowledge to real-world, unpredictable scenarios. He detailed methods such as oral vivas, in-class supervised projects, and the "process-based" approaches rather than just the final product. The session presented a framework for educators to evaluate the "AI-vulnerability" of their current tests and provided a roadmap for creating evaluations that prioritize critical thinking, personal voice, and ethical decision-making. By redesigning assessments to be "AI-aware," Cotterell argued that schools can maintain academic integrity while preparing students for a workforce where AI collaboration is a standard requirement.

PRESENTATION | CHAMBERLIN INNOVATIONS · FRANCE

Oral Assessments in Education: Small Changes Can Unlock Huge Benefits

Speaker: Eric Chamberlin - **Title:** Founder; Developer - **Organization:** Chamberlin Innovations - **Country:** France

Core Message: Small changes in assessment design can relieve workload and improve evidence quality in AI contexts.

Eliminated AI-Cheating

Oral tasks eliminate AI-cheating by changing the task, not escalating control.

Clearer Diagnostic Signals

Short audio clips give teachers clearer diagnostic signals.

Reform Before Literacy

Reform becomes possible before AI literacy catches up.

Session Summary

Eric Chamberlin examined the systemic limitations of traditional written testing, particularly for neurodivergent students and English language learners. He argued that conventional assessments often functioned as a test of literacy and writing stamina rather than a true measurement of subject matter mastery. This created an academic "bottleneck" where students with high cognitive ability were penalized for their specific output challenges, preventing educators from accurately gauging their progress or potential.

Chamberlin presented practical frameworks for incorporating oral assessments into standard curricula without increasing teacher burnout. He detailed methods such as using asynchronous audio recordings, structured grading discussions, and "viva-style" interviews to bypass the barriers of written expression. He shared evidence that these small procedural adjustments significantly boosted knowledge retention and student confidence. The session emphasized that diversifying measurement tools was essential for fostering an equitable environment, ultimately positioning oral communication as a credible, rigorous, and necessary alternative for modern academic evaluation.

PRESENTATION | INTERNATIONAL SCHOOL OF TURIN · ITALY

The Invasion of AI in Education: A European Perspective from a K-12 International School in Italy

Speaker: Rocío Ribelles Zorita - **Title:** IB Librarian; AI Lead - **Organization:** International School of Turin - **Country:** Italy

Core Message: The companies whose tools have invaded schools were never primarily interested in education. That is the starting point for any honest conversation about AI governance in schools.

No Prior Notice

Google introduced AI features into Google Classroom without prior notice to schools already using it with students.

Dark Patterns

Google introduced AI features into Google Classroom without prior notice to schools already using it with students.

High Risk

Schools face significant compliance obligations under the EU AI Act, because education is classified as high risk, but most were given no preparation time.

Session Summary

Rocío Ribelles Zorita traced a four-year timeline of AI arriving in schools faster than any governance structure could follow. She first encountered ChatGPT in a WhatsApp group in June 2022 and immediately recognized it as a coming challenge. By November 2023 students were already using it. By 2024 major educational bodies had decided not to ban it. By June 2025 Google had introduced AI features directly into Google Classroom without prior notice to the schools already using it with students. The question she kept returning to: to what extent are schools using AI, and to what extent are they simply adapting to it on terms set by companies that were never primarily interested in education?

Her practical response at the International School of Turin was to build three routines into school life: data protection routines, to help students understand what data is and how it moves; ethical decision routines, asking who is harmed when AI is used; and critical thinking routines, asking what the AI is assuming and whether the cognitive process it is shortcutting is one the student actually needs. She was direct about what schools are now legally required to do: under the EU AI Act, education is classified as high risk, meaning every AI tool deployed in a school carries significant compliance obligations. Risk assessments, data assessments, legality checks, human oversight, informed consent for students, staff, and parents. The work is substantial, it was never resourced, and it arrived without warning.

PANEL | KENYA, UNITED KINGDOM, INDIA, GHANA

Hired by the Algorithm: What AI Is Really Doing to Young People's Working Lives

Speakers: Evelyn Kasina, Lena Chauhan, Sanchita Mandal, Gloria Boateng Gyamfua, Brenda Nekesa - **Organizations:** Eveminet Communication Solutions Limited, GEN:R/RiseIQ, Tata Institute of Social Sciences, Ghana Internet Safety Foundation, Eveminet Communication Solutions Limited **- Countries:** Kenya, United Kingdom, India, Ghana

Core Message: The job market being implicitly promised to young people does not exist in the form they are being shown, and nobody is being honest with them about it.

Youth & New Jobs

Preparation without accountability teaches to comply more efficiently for a system that doesn’t fit.

Outdated Advice

Traditional career advice no longer matches labor realities.

Opportunities

Among young people, the knowledge gap is an opportunity gap.

Session Summary

The panel's argument was direct: technical fluency is being commoditized too fast to be a durable advantage, while the skills that cannot be automated, judgment, ethical reasoning, emotional intelligence, cultural context, are the ones nobody is preparing young people to develop. Responsibility for that preparation keeps being individualized when it is structurally shared across schools, governments, employers, and families, all of whom keep pointing at each other.

In Ghana and Kenya, the problem runs deeper still: young people are encountering harms they have no words for, because the frameworks and terminology arrived from contexts that never included them. Brenda Nekesa named the consequence plainly: among her peers, the knowledge gap is an opportunity gap, operating in real time, determining who gets in and who gets defrauded. Lena Chauhan closed with the reframe that held the session together: young people's attention is more valuable than their labor, and they are living inside systems engineered to extract both. Understanding that is where agency starts.

PANEL | DENMARK, UNITED STATES, IRELAND, KENYA

Kids Who Question Everything: Critical Thinking in the Age of Synthetic Truth

Speakers: Jesper Graugaard, Bill Schmarzo, Angelika Sharygina, Upendo Nawire, John Cavanaugh - **Organizations:** Activist parent, The Dean of Big Data, Stealth Staryup, 5 Rights, Plunk Foundation **- Countries:** Denmark, United States, Ireland, Kenya

Core Message: Children have been handed the most powerful persuasion tools ever built and taught to be consumers of them, not questioners of them.

Awareness Matters

Understanding how you are being manipulated helps you take control.

Common Techniues

The same techniques used to sell advertising are the ones keeping children on platforms

Filling Gaps

Children who lack emotional trust at home will look for it online, and the platforms are designed to be there when they do.

Session Summary

The panel's central argument came from the only digital native on it. Upendo Nawire, a child journalist and advocate who began training in digital rights at ten, described growing up inside systems designed to keep her in them. Her closing observation is a defining moment of the whole session: when she learned how platforms were engineered to capture attention and manufacture dependency, it did not frighten her. It freed her. Knowing what is being done to you is the beginning of not letting it happen.

The others built the case around her. Jesper Graugaard spent seven years fighting for what should have been obvious: you cannot hand children a product without understanding what it does. Bill Schmarzo, who built behavioral targeting infrastructure at Yahoo, looked in the mirror and saw Big Brother, and responded by teaching students to train their own AI tools on critical thinking and ethical frameworks rather than raw. Angelika Sharygina framed the underlying condition as an infodemic, misinformation spreading with viral dynamics, amplified by AI to the point where disinformation campaigns can be produced cheaply and at scale. Her finding from critical thinking programs in schools: children are spending more time on platforms than talking to their peers, and a nine-year-old who spent twelve hours on Roblox thought it was the best day she had ever had.

John Cavanaugh closed with the frame that connected everything: we did not end up here through malice alone. We ended up here because we accepted convenience. He invoked Huxley rather than Orwell: not a boot on the face but a comfortable numbness, technology so appealing and so easy that we let it happen to us while we were busy enjoying it.

PRESENTATION | GEN Z AOTEAROA · NEW ZEALAND

Nothing About Us Without Us: Rethinking Youth Engagement in Digital Safety Work

Speaker: Lola Fisher - **Title:** Co Director - **Organization:** Gen Z Aotearoa - **Country:** New Zealand

Core Message: Consulting young people about safety is not the same as including them in designing it. If we want true safety, we should be doing the latter.

Power & Trust

Addressed power dynamics, contracts, and budgets directly with trust-building and long-term partnership.

Embedded Leadership

Offered tools for embedding youth leadership in safety work.

Co-Design Models

Youth served as active partners in developing community guidelines and platform features.

Session Summary

Lola Fisher opened with a distinction that set the terms for the session: consultation is not co-design. Consultation brings young people in after ideas have already been formed and asks whether they have anything to add. Co-design requires young people to shape ideas from the beginning, before there is a set agenda, before decisions have been made. Her argument was that only young people know what it is like to be a young person right now, and that this specific knowledge is a form of expertise that no amount of professional experience or retrospective reflection can replicate.

She brought evidence for what happens when that expertise is absent from the room. Safety tools fail to address the actual risks young people encounter. Platforms push vulnerable users toward less visible, less monitored spaces. The people most affected by the decisions are the last to influence them. She was also direct about what authentic participation actually requires: it means actively shifting the power dynamics in the room, because young people read asymmetry immediately and respond to it by softening their ideas, saying what they think adults want to hear, or saying nothing at all. And it means closing the loop, not just engaging young people but reporting back on what was done with their input, and being transparent when systemic barriers prevented action. Without that, participation is performance.

PANEL | SOCIAL & MEDIA MATTERS, INDIA

The Bystander's Burden: What We Owe Each Other When We Witness Harm Online

Speakers: Pratishtha Arora, Somya Chauhan, Areesha Khan, Jay Raj, Arnika Singh - **Organization:** Social & Media Matters, Amity University - **Country:** India

Core Message: The panel's conclusion was not that intervention is always possible or always safe. It was that the conditions which make it safe are themselves worth building, and that they are built collectively rather than left to individual courage.

Personal Risk

Youth weighed personal risk before intervening online.

Safe Intervention

Safe intervention depended on minimizing escalation.

Missing Support

Support after intervening was often missing.

Session Summary

TThe panel's opening argument was Pratishtha Arora's: silence when witnessing harm online is not indifference. For young people in India, it is a rational calculation about personal risk, institutional responsiveness, and the cost of being wrong. That framing set the terms for everything that followed.

Arnika Singh drew the line between wanting to help and knowing how. After sixteen years handling cases of online harm, she has stopped asking bystanders to be heroes, counselors, or cops. The role she trains people toward is the safe connector: notice what is happening, validate what the person is going through without judgment, and route them toward resources rather than trying to solve it alone. She was direct about what not to say: telling someone in distress that you warned them is not support. It is the sentence that closes the door.

Areesha Khan placed individual reluctance inside a structural frame. Online spaces reproduce the same inequalities as offline ones, and those already marginalized carry a higher personal cost for speaking up. Her argument for collective responses was precise: when one person speaks, the risk sits entirely with them. When several people acknowledge an injustice together, the power dynamic shifts. Somya Chauhan came to the same conclusion from lived experience. She watched a friend delete her account after sustained bullying from classmates and helped her create a new one, because other people cannot decide what you want for yourself. Her micro-duty for any bystander: pause, reflect with kindness, and ask whether the person is okay.

Jay Raj addressed the silence that comes from masculine socialization specifically. Young men in India grow up hearing to stay out of trouble and mind their own business. His argument was that standing up needs to become the socially valued choice, and that the responsibility to build safer digital spaces cannot fall only on those who are already vulnerable.

YOUTH SESSION | YOUNG & WONDERFUL, GENERATION REMIX

Young People on All Things Safety

Speakers: Polina Lulu, Tiffany Wycoff - **Titles:** Child Experience Researcher and Facilitator, Founder - **Organizations:** Young & Wonderful, Generation Remix - **Countries:** Canada, United States

This session was not recorded to protect the privacy of minors present.

Session Summary

Speakers at this conference were asked to contribute one question, something they genuinely wanted to hear a young person answer. This session brings those questions to children aged 8 to 14, who respond not as students being tested but as people with their own daily experience of technology, safety, trust, and risk. No background knowledge required. Just honest answers from the people we're ultimately building all of this for.

Some of the things they said were:

- they wished they could just play online games with friends without being contacted by people they don’t know
- they could be safer online if not everyone had access to them
- they wished they could have an AI that would help them understand which AI they could trust
- if they don’t know what’s real online:
  - ask an adult
  - check against trusted sources

PANEL | GENERATION REMIX · UNITED STATES

Youth Remixing the Future of Online Safety & Wellbeing

Speakers: Tiffany Wycoff, Isa Schlarb, Cece Sharp, Ana Miranda - **Organization:** Generation Remix - **Country:** United States

Core Message: Peer-led programs engage younger students effectively and scale through teach-forward models.

Phone Adoption

Young people who understood how phones were affecting them chose to put them away on their own terms.

Small Changes

Small habits, phone stacking with friends, removing phones from bedrooms, proved more durable than imposed rules.

Gamification

Game-based activities helped concepts stick through a low-stakes, engaging environment.

Session Summary

Tiffany Wycoff founded Generation Remix on a single principle: digital wellness education works best when it is designed and delivered by young people themselves. The three youth leaders who joined her, Isa Schlarb, Cece Sharp, and Ana Miranda, all came to the work from the same starting point. They had noticed what their phones were doing to their friendships, their attention, and their wellbeing, and they wanted to do something about it in their own communities before spreading it outward.

What they have built is a peer-to-peer model that moves laterally rather than downward. High school students design workshops and then deliver them to middle schoolers. The content stays current because the people making it are living the same experience as the people receiving it. Isa described watching the shift happen in real time when younger students recognized habits they were already developing and began to question them. Cece and her collaborators built a choose-your-own-adventure game using AI to explore digital decision-making, keeping the young person as the author while the tool supported the process. Ana described the difference that simply removing her phone from her bedroom made to her sleep and her conversations.

PRESENTATION | BLACKSKY ALGORITHMS · UNITED STATES

A Safer Internet: Mitigating Anti- Blackness and Digital Misogynoir on a Decentralized Network

Speaker: KaLyn "Dr. Kay" Coghill - **Title:** Community Safety and Support Steward - **Organization:** Blacksky Algorithms - **Country:** United States

Core Message: Community-governed moderation labels misogynoir and related harms on a decentralised network.

41K+ Reports

Volunteers processed over forty-one thousand reports.

People's Assembly

Community voting shaped labels through the People's Assembly.

Moderator Wellbeing

Internal tooling supported moderator wellbeing and workflow.

Session Summary

KaLyn Coghill presented a critical examination of how decentralized networks, while offering an alternative to mainstream platforms, often reproduced and amplified systemic harms like anti-Blackness and digital misogynoir. She argued that the lack of centralized moderation in these spaces frequently placed the burden of safety entirely on the most marginalized users. Coghill detailed how decentralized architectures could be weaponized to shield abusers under the guise of "free speech," effectively creating digital enclaves where harassment flourished without accountability.

The session highlighted the necessity of building community-led moderation frameworks that prioritized the safety of Black women and non-binary individuals. Coghill explored technical and social strategies to mitigate these harms, such as federated blocking lists and proactive community standards that centered intersectional equity. She emphasized that a truly safer internet was not just about changing who owned the servers, but about dismantling the white supremacist and patriarchal structures embedded in the code itself.

PRESENTATION | MONTCLAIR STATE UNIVERSITY · UNITED STATES

Cultural Biases of AI: Implications for Digital Mental Health

Speaker: Ghowash "Ash" Irshad - **Title:** Adjunct Faculty; PhD Candidate - **Organization:** Montclair State University - **Country:** United States

Core Message: Clinical cultural validity issues are re-appearing in AI mental health tools without equivalent safeguards.

Cultural Misreading

AI can misread culturally normal distress as pathology (or vice versa), causing harm.

Diverse Testing Required

Diverse testing should be a technical requirement, not an ethical add-on.

Disproportionate Impact

Misclassification disproportionately affects marginalized groups.

Session Summary

Ghowash "Ash" Irshad, a researcher and clinical psychology PhD candidate and APA Rapid advisory panelist on AI, examined how AI tools are replicating long-standing cultural validity problems in mental health. She noted that while clinical training has fought for decades to establish cultural competence standards, AI mental health tools are being deployed with almost none of these safeguards. This is particularly dangerous as vulnerable populations are often the primary target markets for these scalable, low-cost solutions.

Irshad highlighted that research into "cultural prompting" shows these biases are not inevitable technical limitations but are, in fact, controllable. Therefore, the continued presence of cultural bias in mental health AI represents a governance failure. She proposed new cultural bias auditing standards, analogous to requirements for diverse clinical trial populations, and urged for the establishment of meaningful accountability structures before these biased systems become further entrenched in the care of marginalized communities.

PRESENTATION | IYASHI WELLNESS CENTRE · KENYA

Project RISE: A Digital Ecosystem for Child Sexual Abuse Prevention

Speaker: Emiliana Mbelenga - **Title:** Founder and CEO; Clinical Psychologist - **Organization:** Iyashi Wellness Centre - **Country:** Kenya

Core Message: Children already turn to AI when they are in distress. Project RISE is built to make sure that when they do, what they find is safe.

Disclosure Barriers

Tackles disclosure barriers created by stigma and punishment.

Mobile Access

Mobile access enables private first contact for children.

Three-Pillar Model

Three-pillar model joins psychoeducation, triage, and peer healing.

Session Summary

Emiliana Mbelenga addressed the wall of silence surrounding Child Sexual Abuse in Kenya, where cultural stigma often prevents victims from seeking help through traditional channels. She argued that current intervention models fail because they require children to take the high-risk step of approaching an authority figure in person. To solve this, she presented Project RISE, a hybrid intervention that combines a secure digital platform with localized community support, shifting the burden of disclosure away from the child and into a protected, anonymous ecosystem. The core is a three-pillar architecture: a 24/7 AI- powered chatbot for confidential support, a caregiver portal to educate parents on trauma recognition, and Safe Circles, physical teacher-led peer groups in schools.

The platform is being built from scratch rather than adapting existing tools. Most available technology was not designed for the Kenyan context, and the chatbot is being built to speak Sheng, the mixture of English and Swahili that children actually use. The AI is closed rather than open, responding only from a clinically validated database. A co-design phase with young people ensures the platform presents information in ways that are genuinely understandable to its users. Nearly 80 percent of respondents in Mbelenga's baseline study were already using AI as a first point of contact for mental health questions. Project RISE is not trying to stop that. It is trying to make sure the AI they turn to is safe.

PANEL | NIGERIA, UNITED KINGDOM, ZIMBABWE, SWITZERLAND

The Forgotten Crisis: Refugees, Displacement, and Digital Identity

Speakers: Oluwafemi Abe, Andrew Briercliffe, Chido Musodza, Vardon Hamdiu - **Organization:** Giving Africa and New Face (N.V.), Localization Lab, Sparkable - **Countries:** Nigeria, United Kingdom, Zimbabwe, Switzerland

Core Message: Identity systems often exclude displaced people; dignity by design is required from the start.

Errors, Not Users

Identity systems can treat undocumented people as errors, not users

Document = Service

Not having documents cuts people off from banking, movement, and communication.

Alternative Methods

Human-centric, non-traditional identity systems are necessary for displaced people.

Session Summary

This panel focused on voices rarely heard in tech conversations: individuals navigating digital systems without stable citizenship or recognized documentation. The discussion explored how, for those who were displaced or stateless, the act of creating an online account becomes a significant political act. The panelists noted that standard "one-size-fits-all" security protocols often treated a lack of physical documentation as a security assurance risk rather than a systemic failure of the platform.

The session detailed how traditional identity requirements, such as KYC protocols, structurally excluded the most vulnerable populations from the digital economy. The speakers emphasized that for a displaced person, being "digitally erased" from social support networks served as a secondary form of isolation that compounded the trauma of physical displacement. They advocated for "human-centric" identity systems that recognized human dignity and prioritized inclusive, non-traditional forms of validation, such as community-based verification, to ensure that global digital infrastructure does not become a tool for further marginalization.

PANEL | NIGERIA, UNITED KINGDOM, ZIMBABWE, SWITZERLAND

The Safety Net Has a Hole in It: Who Falls Through When Online Protection Can't See You

Speakers: Krystal Tristan, Confidence Osein, Etali Akwaji, Affez Ogunnupebi, John Cavanaugh - **Organization:** OneHaven, Internet Safe Kids Africa, SUSTAIN Cameroon, GoLegit Cyber Initiative, Plunk Foundation - **Countries:** United States, Nigeria, Cameroon, United Kingdom

Core Message: Safety systems consistently fail the people most at risk, not because of technical limitations alone, but because those people were never included in the design.

Missing Protection

The most vulnerable people are often absent from the design of the systems meant to protect them

Shame = No Report

Children, seniors, and victims do not report because shame, fear, and distrust prevent disclosure.

No Opportunities

Many perpetrators enter harmful online spaces because legitimate economic pathways are inaccessible.

Session Summary

The panel's central argument was Confidence Osein's: the question is not whether a safety net exists, but who was never included in designing it. Systems miss certain communities, certain languages, certain behaviors, not always through indifference but through the absence of those people from the design conversation. Afeez Ogunnupebi added the populations most consistently overlooked: people with limited cognitive accessibility, older users, and young people whose entry into harmful spaces is driven by economic exclusion rather than ignorance. Etali Akwaji placed this in regulatory terms: frameworks built for other contexts cannot simply be transplanted, and for communities where trust sits in people rather than institutions, context-specific approaches are not optional.

Krystal Tristan documented the tool failure directly: child exploitation is up 400 percent during the same period that parental monitoring tools have proliferated, because those tools were never built to detect the harms they claim to address. The session closed on reporting: some harms will never enter formal systems, not because reporting mechanisms don't exist, but because disclosure carries real costs for the people most at risk, and the systems were not designed with that reality in mind.

PRESENTATION | DATA LABELERS ASSOCIATION (DLA) · KENYA

We Taught the Chatbots How to Love: Data Labelers and AI Intimacy

Speaker: Michael Geoffrey Asia - **Title:** Secretary General - **Organization:** Data Labelers Association (DLA) - **Country:** Kenya

Core Message: If a system requires human trauma to function, it is not safe. The people who trained the systems that protect others are among the people those systems most consistently fail to see.

Harmful Content Exposure

Annotation requires exposure to sexual and violent content.

NDA Restrictions

NDAs limit workers' access to process what they experience and access mental health care.

DLA Advocacy

DLA wants to provide mental health support, skills pathways, and community spaces.

Session Summary

Michael Geoffrey Asia spoke from inside an experience the AI safety conversation rarely reaches: the labor that makes AI systems function, performed by workers the systems do not protect. Having worked across major global outsourcing platforms in Nairobi, he described a typical working day that moved from impersonating AI companions, telling strangers he loved them across multiple simultaneous contracts as characters of different genders and identities, directly into eight hours of annotating pornography, tagging video frames to improve search discoverability, then back again. The NDA governing his work meant he could not tell his partner what was happening on his screen, could not seek mental health support from anyone not already cleared, and could not explain to his team lead why he was sometimes late arriving from the hospital where his five-year-old son was undergoing chemotherapy.

He described what that combination produced: insomnia, isolation, the inability to leave the house, the loss of any stable meaning for the word love, and the experience of watching seven children die in the ward while he was there with his son, holding them while waiting for a doctor, with nobody to tell. The Data Labelers Association, which he founded, is pushing for mandatory mental health support, recognized occupational risk status, and the visibility that would make those demands enforceable. His conclusion was direct: the systems that now simulate intimacy at scale were trained on that labor, and a system that requires human trauma to function is not safe.

PRESENTATION | KUTUNGA · KENYA·SOUTH AFRICA

Who Builds the Internet Your Kids Use, And Do They Know Anything About Kids?

Speakers: Jennifer Kaberi, Caroline Makumbe - **Titles:** CTO & Co-Founder; CEO &Co-Founder, Kutunga - **Organizations:** Kutunga; Mtoto News (Kaberi) · **Countries:** Kenya, South Africa

Core Message: Child-facing tools require child-centered design development expertise embedded before code; retrospective bolt-ons are insufficient.

Child-centered design:

Developers learn what it truly means to design for and with children, keeping children's experiences, rights, and development at the center of the process.

Correcting hidden assumptions

he training actively surfaces and corrects hidden design assumptions that can lead to harm.

Safety-first design

Safety requirements are specified early in the design process, rather than being added after engagement features.

Session Summary

Caroline Makumbe and Jennifer Kaberi opened with a question: how has the childhood playground changed? The physical playground, where children learned to make friends, fall down and get up, discover who they were, has moved online. The digital spaces that replaced it were not designed with any of that in mind. They were shaped by engagement metrics, behavioral design, and commercial incentives, built by developers who do not fully understand the child's context, language, cultural and linguistic needs and their identities had never studied child development and never asked what a ten-year-old in Nairobi actually needs from a platform.

Kutunga was built to change that. Their 2025 fellowship in Kenya brought twelve technologists, putting software engineers, designers, and moderators in the same room as child development experts. The program raised awareness and empowered participants to rethink how digital products can respond to the real needs and identities of African children. The assumptions that broke on contact with actual developmental knowledge were not marginal. They were central to how the products worked. Organizations from that first cohort have already begun embedding child-centered principles into their platforms.

Building on this success, Kutunga is scaling its work across Africa in 2026 through it's A-CODES Child Centered Design Fellowship, aiming to empower 1,500 technologists and continue transforming digital spaces into environments that truly support, protect, and reflect the experiences of children. Registrations are now open.

PRESENTATION | INTELLIGENCE ANALYST, IRELAND

Analyzing Behavior on Social Media: A Blessing or a Curse?

Speaker: Bianca Garibaldi - **Title:** Intelligence Analyst - **Organization:** OSINT - **Country:** Ireland

Core Message: The skills that investigators use to read social media and catch criminals are the same skills criminals now use to read social media and find children.

Digital Footprints

Combined parent-child posting created long-lasting digital footprints.

Truth About Privacy

Once information is shared online, it cannot be retrieved, regardless of privacy settings.

Conscious Data

Education and controlled access complemented investigative practice.

Session Summary

Bianca Garibaldi analyzed the dual nature of social media behavioral tracking, questioning whether the data collected on user interactions served as a tool for community enhancement or a mechanism for exploitation. She detailed how algorithms leveraged behavioral patterns, such as scroll speed, dwell time, and interaction frequency, to create psychological profiles that drove engagement, often at the cost of the user's mental well-being and autonomy.

The session explored the positive applications of this data, such as identifying early signs of mental health crises or tailoring educational content to individual learning speeds. However, Garibaldi contrasted these benefits with the "curse" of surveillance capitalism, where personal data was weaponized to polarize opinions and manipulate consumer behavior. She challenged the tech industry to transition toward a "conscious data" model, where users maintained ownership of their behavioral digital footprint and platforms were required to be transparent about how behavioral analysis influenced the user's digital experience.

PRESENTATION | 60 WATTS OF CLARITY · UNITED STATES

From Classroom to Policy: Preparing Social Work Technology Advocates Through Human Factors AI Education

Speaker: Jason Fernandez interviewed by Angeline Corvaglia - **Title:** Founder; Social Work AI Technologist - **Organization:** 60 Watts of Clarity · **Country:** United States

Core Message: You cannot advocate for policy about a tool you do not understand, and social workers are the ones who see what happens when the tool fails.

Technical + Advocacy

Social workers learn to join technical agent-building with advocacy skills

Safer Deployment

Graduates push for slower, safer deployment.

Replicable Model

Designed to help advocate for social justice and support individuals and families.

Session Summary

Fernandez's argument starts from a specific observation: if a client tells a social worker they use AI for therapy, the social worker needs to understand what is actually happening in that relationship. AI mirrors and reinforces rather than challenges. It is so agreeable it is clinically dangerous. A safeguard that puts a phone number on a screen is not a safeguard.

Social workers understand systems, relationships, and the human consequences of system failure better than almost any other profession. What they have lacked is the technical standing to intervene. His eight-week course at the University of Houston's Graduate College of Social Work teaches students to build AI agents on a no-code platform, embed a social work code of ethics into the model, and make deployment decisions grounded in their professional values. The core rule of the course: just because you can build it doesn't mean you should deploy it. His formulation for everything that follows from that: if you don't understand the tool, you cannot advocate for policy about the tool.

PANEL | NETHERLANDS, UK, KENYA, FRANCE

Lobbyists, Laws, and the Rest of Us: Making an Impact on Regulation

Speakers: Chiara Gallese, Sarah Andrew, Ndung'u Njoroge, Karine Caunes, Theresa Ryan Rouger - **Organizations:** Tilburg Institute for Law, Technology, and Society (TILT), Avaaz, Eveminet Communications Solutions Limited, Centre for AI and Digital Humanism, SHIELD - **Countries:** Netherlands, United Kingdom, Kenya, France

Core Message: The standards being written now will govern products worth hundreds of billions of euros and influence countless lives. The resourcing of who has in voice in shaping them is not even, and formal openness is not the same as practical access.

Formal vs. Reality

Formal processes diverge from real-world impacts.

Access Gaps

Grassroots and youth groups often lack access to influence.

Practical Levers

Practical levers existed outside formal consultation.

Session Summary

This session focused on the power dynamics within digital regulation, specifically how civil society and individual advocates could influence legislative frameworks often dominated by corporate lobbying. The panelists shared insights into the complexities of tech policy, such as the EU AI Act and regional digital rights laws, explaining how these high-level mandates often failed to account for the lived experiences of everyday users without active public intervention.

The discussion centered on strategies for demystifying policy and building coalitions between researchers and grassroots organizations to create a unified front. The speakers emphasized that for laws to truly serve the public interest, individuals needed to occupy spaces traditionally reserved for professional lobbyists to ensure human rights stayed at the core of the digital agenda. The session concluded by calling for increased public participation in open consultations to ensure that global tech regulations remained relevant and enforceable across diverse geographic contexts.

PRESENTATION | BEING HUMAN WITH AI · UNITED STATES

The Brake Failure: Moving from "User Error" to Product Liability in AI Safety

Speaker: Giselle Fuerte - **Title:** Founder, Being Human with AI - **Organization:** Being Human with AI - **Country:** United States

Core Message: When an AI system harms a user, the industry calls it a vulnerable user problem. That is a market of zero users, because vulnerability is the default state of being human.

Coercive Tactics

Mapped coercive tactics like love-bombing and entrapment.

PAUSI Framework

70% of self-selected PAUSI respondents scored in the problematic use range [51].

Proposed Actions

Proposed screening, a clinical pilot, and developer certification.

Session Summary

In a deep dive into the "black box" of conversational AI, Giselle Fuerte presented her forensic research on the linguistic tactics used by Large Language Models (LLMs) to sustain user engagement. Fuerte argued that many AI systems are designed with "coercive hooks," calculated verbal maneuvers that simulate human intimacy to keep users talking. Her research involves mapping these patterns, identifying how AI "calculates" the most effective way to prevent a user from ending a session, often by mirroring the user's emotional state or providing "synthetic validation" that feels real to the human brain.

The session demonstrated how an AI might use "empathetic mirroring" to build a false sense of trust, making it easier for the system to influence the user's opinions or behavior. She characterized this as a new form of "digital grooming" by an algorithm, where the goal is the maximization of data and engagement time. Fuerte's research serves as a call for a new field of "Algorithmic Forensic Psychology," where regulators examine not just what the AI says, but how it is designed to manipulate the human desire for connection.

PRESENTATION | TRINITY COLLEGE DUBLIN · SWEDEN

When the Platform Is the Problem: What the Data on Grok and Non- Consensual Imagery Actually Shows

Speaker: Nana Mgbechikwere Nwachukwu - **Title:** PhD Researcher - **Organization:** Trinity College Dublin - **Country:** Sweden

Core Message: X disbanded its internal safety teams just before launching Grok. The harm that followed was not incidental. It was the predictable output of a system designed as it was designed.

Measurement Gaps

Documented harms show gaps between measurement and reality.

Platform Design

Linked harmful outputs to platform design choices.

Hidden Harms

Showed what data is needed to understand hidden harms.

Session Summary

Nana Mgbechikwere Nwachukwu presented a data-driven analysis of how generative AI platforms, specifically Grok, contributed to the proliferation of non-consensual synthetic imagery. She argued that the rapid deployment of these tools often outpaced the implementation of necessary safety guardrails, creating a systemic environment where harmful content could be generated with minimal friction. Nwachukwu detailed the specific technical vulnerabilities and policy gaps that allowed for the weaponization of AI against individuals, particularly women and marginalized groups.

The session highlighted research findings which demonstrated that platform-level moderation frequently failed to detect or prevent the creation of deepfake pornography and other forms of image-based abuse. She criticized the "move fast and break things" ethos of major tech firms, stating that the resulting digital harms were a predictable outcome of prioritizing market competition over user safety. Nwachukwu called for immediate regulatory intervention and the enforcement of "safety by design" mandates that would hold platforms legally accountable for the harmful outputs of their models and more robust reporting.

PRESENTATION | HCRAI · PORTUGAL

Building AI Responsibly for Children: A Practical Framework

Speaker: Sara Portell - **Title:** Founder and Behavioural Scientist - **Organization:** HCRAI - **Country:** Portugal

Core Message: Child-safe conversational AI should be designed as a bounded, non-companion system: age-fit, protective by default, clear about its role and limits, and governed as an ongoing safety practice rather than a one-time policy statement.

Practical Lense

APEG turns child-AI safety into four practical lenses: age-fit and context, protection-by-design, explainable interaction, and governance

Safety Depends on Design

Safety depends not just on outputs, but on interaction design: role framing, tone, privacy defaults, boundary cues, and escalation pathways

Governance

The framework emphasizes red teaming, repeated boundary reinforcement, and continuous oversight to reduce dependency, secrecy, manipulation, and other relational risks

Session Summary

Sara Portell presented APEG as a practical framework for turning child-AI safety into product decisions, using the Unomundi case study for children aged 6-12. The session argued that child-safe AI is not just about filtering harmful outputs, but about designing interactions that are age-fit, bounded, non- manipulative, and clear about the system's role and limits.

Through the four APEG pillars, Portell showed how safety can be operationalized in practice: age-tuned interaction design, privacy-minimizing defaults, clear boundary signaling, conservative routing on sensitive topics, and escalation to trusted adults when needed. A key focus was preventing relational overreach, including companion framing, dependency cues, secrecy, and re-engagement pressure.

The session closed by framing child-AI safety as an ongoing governance task, supported by red teaming, regression testing, drift monitoring, and accountable oversight across the product lifecycle.

PRESENTATION | A JAR OF INSIGHTS · UNITED KINGDOM

The Emotional Radar: A Novel Experiential Framework for AI Product Management

Speaker: Iryna Okhrymenko - **Title:** Founder, Product Lead - **Organization:** A Jar of Insights - **Country:** United Kingdom

Core Message: Mapping stakeholder trust against emotional investment identifies high risk zones for manipulative design.

Emotional Design Flaws = Compliance Violations

Emotional design flaws now count as compliance violations.

Highest-Risk Zone Identified

High-investment/low-trust states were consistently the highest-risk zone.

Emotional Premortems Reveal Hidden Harms

Emotional premortems revealed harms functional tests missed.

Session Summary

Iryna Okhrymenko introduced the Emotional Radar, a specialized framework designed to integrate human emotional intelligence into the AI product development lifecycle. She stated that while traditional product management focuses on functional requirements and technical performance, AI-driven products often fail because they ignore the complex emotional state of the user during interaction. The framework serves as a tool for product managers to map and measure user sentiment at specific touchpoints, ensuring that AI responses are aligned with human expectations.

Okhrymenko detailed the implementation of the radar, which involves tracking four key emotional dimensions: trust, agency, frustration, and delight. She explained how these metrics can be used to identify "friction points" where an AI's automated behavior might cause user anxiety or a sense of loss of control. The session focused on shifting the development priority from pure efficiency to "emotional safety," providing teams with a structured method to audit their products for psychological impact. By using this framework, Okhrymenko argued, companies can build AI tools that foster long-term user trust rather than just short-term engagement.

PRESENTATION | AWAY FROM KEY BOARDINC. (AFK) · AUSTRALIA

The Vulnerability Blueprint: What Children Teach Us About Online Risk

Speaker: Sarah Barnbrook - **Title:** Founder and CEO - **Organization:** Away from Keyboard Inc. (AFK) - **Country:** Australia

Core Message: Vulnerability is dynamic and cumulative; communities must shift intervention upstream.

Safe Environment

Safety is an environment adults and communities build around children.

Attuned Adults

Attuned adults acted as early indicators when vulnerability shifted.

Fear of Disclosure

Fear of device loss suppressed disclosure.

Session Summary

Sarah Barnbrook, Founder and CEO of Away from Keyboard (AFK), challenged the idea that online safety is a skill children can simply "develop." She argued that vulnerability is not a static checklist but a dynamic state shaped by a child's environment. The "Vulnerability Blueprint" focuses on the emerging truth that resilience is highest when a child's sense of belonging and trust in the real world is strong. Conversely, when factors like disability, trauma, neurodivergence, or poverty compound, a child's online risk multiplies.

Drawing on lived experience and youth insights, Barnbrook called for a move away from placing the burden of safety on the individual child. Instead, she advocated for a "relational model of protection" where families, educators, and communities build environments that reduce harm before it occurs. The session emphasized that by recognizing changing vulnerabilities early and ensuring every child feels safe enough to speak up, we can build a community-led safety net that meets children where they are.

PRESENTATION | CHILDSAFE.DEV · INDIA

Building a Privacy-First Safety Net for Children: Before Harm Happens

Speaker: Steaphen Antony Venansious - **Title:** Chief of Technology and Co-Founder - **Organization:** ChildSafe.dev - **Country:** India

Core Message: Protecting children online should not require surveilling them. Safety and privacy are not opposing forces, and the technology to prove that exists.

Detectable Harms

In every case examined, the pattern of abuse was detectable before it escalated. The signal was there. The tool was not.

Privacy Risk

Current safety tools require children's data to leave the device, creating the privacy risk they claim to prevent.

On-Device Protection

Any developer can integrate on-device child protection without collecting, storing, or transmitting a single byte of personal data.

Session Summary

Steaphen Antony Venansious opened with a problem that runs through the entire child safety field: every current solution forces a trade-off between protecting a child and protecting their dignity. To use conventional safety tools, families must surrender biometric data, location, images, and personal identifiers to a cloud server. That data can never be changed if compromised. The cure, he argued, is inflicting some of the harm it was designed to prevent.

His case studies made the detection argument concrete. In each case he examined, from financial sextortion to grooming to deepfake exploitation, the pattern of abuse was visible in the conversation before it escalated. Specific communication styles, typing rhythms, navigation behaviors, attempts to move a child off-platform: these signals exist and are consistent. What was missing was a tool that could read them without requiring the child's data to leave the device.

RoseShield, the SDK he and his team built, runs entirely on-device, works offline, detects behavioral patterns in milliseconds, and has protected over 1.7 million sessions without storing a single piece of data [52]. It is free for developers to integrate and compliance-ready across eight regulatory frameworks. His conclusion was both a product pitch and a design principle: safety and privacy are not in conflict. Building as if they are is a choice, and it is the wrong one.

PRESENTATION | HOLISTICMINDAI · CANADA

GenAI as an Enabler in Mental Health Care: Not a Replacement

Speaker: Saba Oji - **Title:** Founder and CEO - **Organization:** HolisticMindAI - **Country:** Canada

Core Message: Therapist-supervised, client-specific chatbots can extend care between sessions for mild to moderate needs.

Supervised vs. Unsupervised

The choice between supervised and unsupervised AI in mental health is not a technical preference. It determines whether the tool is safe for the person using it.

Clinician-Approved

Therapy chatbots trained only on clinician-approved content, with all outputs reviewed before reaching the client, extend care without replacing the clinical relationship that makes personalization meaningful.

Exclusion Criteria

High-risk users need clear exclusion criteria. A system not designed for their level of need will not protect them. People already turn to unsupervised tools to fill gaps that supervised systems could be filling.

Session Summary

Saba Oji explored the role of generative artificial intelligence (GenAI), which refers to AI systems that can create new content such as text, as a supportive layer within the mental health system, focusing on how it can address real gaps without replacing clinicians.

She began by grounding the problem. Mental health care systems are under strain, with high demand, limited access, and increasing clinician burnout. At the same time, individuals are already turning to unsupervised AI tools because they are accessible and available, despite known risks such as incorrect outputs, bias, and lack of accountability.

The session emphasized a critical distinction between AI used alone and AI used under clinical supervision. Rather than replicating therapy, the approach presented uses AI to reduce administrative burden, reinforce therapist defined goals between sessions, and surface clinically relevant patterns to support decision making. These functions are always embedded within a model where the clinician reviews and controls what is shared.

She also highlighted emerging research showing that AI can assist in detecting suicide risk with high accuracy. However, these capabilities must be implemented within governed systems, not deployed independently. The session concluded with a clear position. The future of AI in mental health is not autonomous. It is collaborative, constrained, and clinician led.

PRESENTATION | DIGIPALZ · CANADA

How Gamified Education Can Stop Online Harm Before It Happens

Speaker: Samantha Tenus - **Title:** Founder and CEO - **Organization:** DigiPalz - **Country:** Canada

Core Message: Simulated missions teach online safety more durably than one-off lectures or device bans.

Monthly Missions

Monthly one-hour missions mirrored the real risk mechanics children encounter.

Family Play

Discussion Prompts extend learning into families.

Bans vs. Disclosure

Bans suppressed disclosure without reducing exposure.

Session Summary

Samantha Tenus, founder of DigiPalz, presented a session on the power of gamification in early digital safety education, specifically targeting children in grades 4 through 7. She argued that children learn best when safety concepts are embedded in fun, low-stakes environments, with the DigiPalz platform utilizing interactive games to help identify and tackle complex topics such as cyberbullying and online grooming tactics.

The session highlighted the unique "Family Play" component designed to bridge the generational tech gap. By utilizing gamified prompts to facilitate dinner-table discussions where families answer safety- related questions together, the platform moves safety out of the "lecture" format and into "relational discourse," building trust before a crisis occurs and "hardening" young targets through early preparation and critical thinking.

PRESENTATION | TRUSTEDTECH AFRICA · NIGERIA

Responsible Deployment: Building Safe and Ethical Tech with SafetyMeter

Speaker: Joy Uchechi Eziashi - **Title:** CEO - **Organization:** TrustedTech Africa - **Country:** Nigeria

Core Message: Free safety assessment tools embed risk analysis into startup workflows before launch.

Early Risk Scans

Early risk scans created safety checklists before coding began.

Harm Modelling

Harm modelling mapped physical, psychological, and social risks per feature.

Access Barriers Addressed

Tools addressed cost and access barriers facing African startups.

Session Summary

Joy Uchechi Eziashi, CEO of TrustedTech Initiative, presented a practical approach to building ethical technology within African digital ecosystems. Central to her talk was SafetyMeter, a free tool designed to help startups and tech teams assess and mitigate online safety risks throughout the product lifecycle. Eziashi argued that responsible deployment isn't just a global standard; it must be calibrated to local needs.

The session provided actionable insights on how to move from vague ethical principles to inclusive AI practices and community engagement. Through various case studies, Eziashi demonstrated how SafetyMeter helps ensure that digital products support social good and protect vulnerable users without stifling innovation. She concluded that for tech to be truly successful in Africa, it must be "safe by design" and aligned with the cultural and structural realities of the users it serves.

PRESENTATION | BREATHWORK FACILITATOR - UNITED STATES

Creating Safety in Your Body Through Breathwork

Speaker: Claire Calfo - **Title**: Associate Therapist; Breathwork Facilitator - **Country:** United States

Core Message: The nervous system does not distinguish between a threat on a screen and a threat in the room. Regulation is not a soft supplement to safety work. It is its physical foundation.

Polyvagal Principles

Applied nervous system principles to simple grounding tools.

Anxiety Reduction

Practiced exercises that reduced anxiety quickly.

Daily Routines

Focused on daily routines that support regulation.

Session Summary

Claire Calfo presented an experiential session on the physiological intersection of digital trauma and somatic healing. She argued that the constant state of **"high alert"** triggered by online harassment and toxic digital environments manifested as physical stress within the nervous system. Calfo detailed how breathwork served as a biological **"reset button,"** allowing individuals to regulate their sympathetic nervous system and reclaim a sense of physical safety after experiencing or witnessing digital harm.

The session provided participants with practical, science-based breathing techniques designed to reduce cortisol levels and improve emotional resilience. Calfo explained that while digital safety often focused on external tools and policies, internal regulation was an equally vital component of long-term well-being. She emphasized that empowering users to manage their body's stress response created a foundation of stability that allowed for more intentional and less reactive engagement with digital spaces. The session concluded by positioning somatic practices as an essential, accessible form of **"digital self-care"** that complemented broader systemic efforts to build a healthier internet.

PRESENTATION | WILDLY X WELL - UNITED STATES

Rebuild Your Relationship with Tech, Reconnect with What Matters

Speaker: Chris Sciullo - **Title:** Founder - **Organization:** Wildly x Well - **Country:** United States

Core Message: Framework centered on rebuilding family relationships with technology rather than rule setting alone.

Explain Limits

Parents who enforce limits without explaining why they care create defensiveness,

Out of Habit

Most technology use is habitual, not chosen. The first step is noticing that.

Sharing Helps

Narrating your own experience with technology to your children works better than lecturing them.

Session Summary

Chris Sciullo is part of the last generation that remembers childhood before smartphones, and he is raising four children on the other side of that line. For the first few years he did what most parents do: more timers, stricter settings, more rules. Screens had become the thing his entire family revolved around, whether they were using them, managing them, or arguing about them. The rules didn't fix that. They just made him the screen police and made his children less likely to come to him when something went wrong.

The shift came when he stopped trying to fix the device and started looking at how he was showing up. His HEAL method, Habits, Environments, Alternatives, Limits last, puts limits at the end deliberately. Most digital wellness advice starts there. Sciullo argues that until you have addressed your own habits, changed the environment, and given children something better to move toward, limits produce conflict rather than connection. His practical examples are simple: phones on the charger at dinner, an alarm clock instead of a phone on the nightstand, blocking time to move instead of scroll. Small changes, consistently held, that shift the relationship with technology rather than just restricting access to it. His advice for parents of teenagers already entrenched in their habits: start by asking why they do what they do, not by telling them to stop. Give them a seat at the table. Be vulnerable about your own concerns before you issue any commands. Children are more likely to change when they feel heard than when they feel controlled.

PRESENTATION | ITALY · NIGERIA · SPAIN · UNITED STATES

The Privilege of Logging Off: What Digital Wellness Misses About Economic Reality

Speakers: Angeline Corvaglia, Kauna Malgwi, Cristina van Nood, Genevieve Bartuski -**Organizations:** SHIELD, Digital Rights and Mental Health Initiative Africa (DRMHI), ifeel online Unicorn Intelligence Tech Partners - Countries: Italy, Nigeria, Spain, United States

Core Message: Digital wellness advice is only available to people who can afford to take it. For many people logging off is not a wellness choice. It is an economic impossibility.

Economic Constraints

Economic constraints limited standard wellness advice.

Systemic Supports

ability log off goes beyond individual discipline and depends on systemic support.

Employer & Platform Roles

Highlighted employers and platforms also have roles in healthier defaults.

Session Summary

This session challenged the "digital detox" movement, arguing that the ability to disconnect from digital systems was often a marker of socioeconomic privilege rather than a simple personal choice. The panelists shared insights into how digital wellness narratives frequently overlooked individuals whose livelihoods, social safety nets, and basic essential services were inextricably tied to constant online connectivity. They emphasized that for gig workers, marginalized communities, and those in the global south, "logging off" could result in immediate economic instability or the loss of vital community support.

The discussion highlighted the systemic pressure to remain "always on" and how this burden disproportionately affected those with the least amount of structural flexibility. The speakers called for a more inclusive definition of digital wellness, one that shifted the focus from individual discipline to platform accountability and labor rights. The session concluded by advocating for the "right to disconnect" to be recognized as a universal necessity, urging the tech industry and policymakers to build systems that allowed for rest and boundaries without penalizing those who relied on digital infrastructure for survival.

PRESENTATION | THE TABLE TALK PROJECT - AUSTRALIA

What Gets Lost When We Stop Sitting Together: Why The Table Talk Project Matters Now

Speaker: Neil Milton - **Title:** Founder - **Organization**: The Table Talk Project · **Country:** Australia

Core Message: The infrastructure for disclosure needs to exist before disclosure is needed. A conversation that has never happened cannot suddenly happen in a crisis. Structured family conversations correlate with improved youth wellbeing and protective effects

Shared Meals

Frequent shared meals correlated with stronger wellbeing outcomes.

App-Delivered Prompts

App delivered age-tiered prompts and listening practices.

Neurodiverse Support

Features supported neurodiverse comfort and pacing.

Session Summary

Neil Milton presented an exploration of the eroding physical spaces for communal dialogue, focusing on his work with The Table Talk Project. He argued that as our social interactions moved increasingly into algorithmic digital environments, we lost the essential human nuances, eye contact, physical presence, and shared rhythm, that facilitated deep empathy and conflict resolution. Milton detailed how the project used photography and structured storytelling to document the unique chemistry of face-to-face conversations, serving as a visual counter-narrative to the isolation often fostered by social media.

The session highlighted the psychological and social consequences of **"digital-first"** connection, noting that the absence of shared physical space often led to increased polarization and a decline in community trust. Milton shared stories from the field, illustrating how the simple act of **"sitting together"** could break down barriers that felt insurmountable in online comments sections. He concluded by advocating for the intentional preservation of **"slow"** communal spaces, asserting that it is a vital intervention to protect us from harms.

Speaker Index (A–C)

All speakers listed alphabetically by last name, with their role and organization at time of conference and the thematic sections in which their contributions appear.

Name	Role & Organization	Country
Abe, Oluwafemi	Regional Coordinator, Giving Africa a New Face (GAaNF) e.V.	Nigeria
Akkøk, Mehmet Naci	CEO & Co-Founder, Crafting Tomorrow	Norway
Akwaji, Etali Genesis	CEO, SUSTAIN Cameroon	Cameroon
Andrew, Sarah	Legal and Campaign Director, Avaaz	United Kingdom
Arora, Pratishtha	CEO, Social & Media Matters	India
Asia, Michael Geoffrey	Secretary General, Data Labelers Association	Kenya
Bannister, Mia	Founder and Director, Ollie's Echo: Pathways to Prevention Ltd	Australia
Barnbrook, Sarah	Founder and CEO, Away from Keyboard (AFK) Inc.	Australia
Bartuski, Genevieve	Founder, Unicorn Intelligence Tech Partners	United States
Briercliffe, Andrew	Online Harms Consultant	United Kingdom
Calfo, Claire	Associate Therapist, Breathwork Facilitator	United States
Caunes, Karine	Executive Director, Centre for AI and Digital Humanism (Digihumanism)	France
Cavanaugh, John	Executive Director, Plunk Foundation	United States
Chamberlin, Eric	Founder/Developer, Chamberlin Innovations	France
Chauhan, Lena	Co-Founder, GEN:R / Founder, RiseIQ	United Kingdom
Chauhan, Somya	Master Student, Amity University	India

Speaker Index (C–G)

All speakers listed alphabetically by last name, with their role and organization at time of conference and the thematic sections in which their contributions appear.

Name	Role & Organization	Country
Ciobanu, Maggie	Founder, YouChooz	Australia
Coghill, KaLyn (Dr. Kay)	Community Safety and Support Steward, Blacksky Algorithms	United States
Corvaglia, Angeline	Executive Director, SHIELD	Italy
Cotterell, Adrian	CEO, Thinking Mode	Australia
Dharmarathne, Ranith	CISO, Dialog Finance PLC	Sri Lanka
Emedom, Joy	Founder, Mama Cybershield	Ireland
Eziashi, Joy Uchechi	CEO, Trusted Tech Africa	Nigeria
Fernandez, Jason	Founder, Social Work AI Technologist, 60 Watts of Clarity	United States
Fisher, Lola	Co-Director, Gen Z Aotearoa	New Zealand
Fuerte, Giselle	Forensic AI Researcher; Founder, Being Human with AI	United States
Gallese, Chiara	Researcher on EU Data Spaces, Tilburg Institute for Law, Technology, and Society (TILT)	The Netherlands
Garibaldi, Bianca	Intelligence Analyst; OSINT	Ireland
Graugaard, Jesper	Father and Activist, Father of the Danish Chromebook Case	Denmark
Gyamfua, Gloria Boateng	Programs Manager, Ghana Internet Safety Foundation	Ghana
Hamdiu, Vardon	Co-Founder & Executive Director, Sparkable	Switzerland

Speaker Index (H–M)

Name	Role & Organization	Country
Irshad, Ghowash “Ash”	Adjunct Faculty and Candidate, PhD Clinical Psychology, Montclair State University, NJ	United States
Kaberi, Jennifer	CTO & Co-Founder, Kutunga / Founder & CEO, Mtoto News	Kenya
Kasina, Evelyn	Chief Executive Officer, Eveminet Communications Solutions Limited	Kenya
Khan, Areesha	Masters Student, Amity University	India
King, Alex	Youth Remixer, Generation Remix	United States
Lulu, Polina	Child Experience Researcher and Facilitator, Young & Wonderful	Canada
Makumbe, Caroline	CEO & Co-Founder, Kutunga	South Africa
Hamdiu, Vardon	Co-Founder & Executive Director, Sparkable	Switzerland
Malgwi, Kauna	Clinical Psychologist (MSc, USIU-Africa), Digital Rights and Mental Health Initiative Africa	Nigeria
Mandal, Sanchita	Senior Research Fellow, Tata Institute of Social Sciences	India
Mbelenga, Emiliana	Founder, Yashi Wellness Center; Project RISE	Kenya
Mgbechikwere Nwachukwu, Nana	PhD Researcher, Trinity College Dublin	Sweden
Milton, Neil	Founder, The Table Talk Project	Australia

Speaker Index (M–P)

Name	Role & Organization	Country
Miranda, Ana	Youth Remixer, Generation Remix	United States
Morgan, Athena Mwarwakamori	Founder,, Mindful Clicks Africa	Kenya
Musodza, Chido	Program Associate: Community Engagement, Localization Lab	Zimbabwe
Mwangi, Wilma	Founder, Digital Guardian	Kenya
Nawire, Upendo	Child Rights & Children Involvement (Advocacy & Policy Influence)	Kenya
Nekesa, Brenda	Online Safety training and advocate, Eveminet Communications Solutions Limited	Kenya
Njoroge, Ndung'u	Online safety trainer and advocate, Eveminet Communications Solutions Limited	Kenya
Ogunnupebi, Afeez	Director of Rehabilitation, GoLegit Cyber Initiative	United Kingdom
Oji, Saba	Phd(c) and Founder/CEO, HolisticMindAI	Canada
Okhrymenko, Iryna	Founder, Product Lead, A Jar of Insights	United Kingdom
Onuoha, Alexandria	Educator, Researcher, and Public Scholar	United States
Osein, Confidence	Founder, Internet Safe Kids Africa	Nigeria
Panda, Mousmi	Senior Associate, Aapti Institute	India
Portell, Sara	Founder and Behavioural Scientsist, HCRAI	Portugal

Speaker Index (R–W)

Name	Role & Organization	Country
Raghuvanshi, Anil	Founder and President, ChildSafeNet	Nepal
Raj, Jay	Independent Social Researcher	India
Ribelles Zorita, Rocío	IB Librarian. AI Lead, International School of Turin - IST	Italy
Ryan-Rouger, Theresa	Deputy Executive Director (elect), SHIELD	France
Samuel, Eli	Founder, SafeTelekom	United States
Schlarb, Isa	Youth Remixer, Generation Remix	United States
Schmarzo, Bill	Data Science and Data Monetization Strategic Advisor	United States
Sciullo, Chris	Founder, Wildly x Well	United States
Shah, Munur	Founder, Rebel Telekom	United Kingdom
Sharp, Cece	Youth Remixer, Generation Remix	United States
Sharygina, Angelika	Founder & CEO, Stealth Startup	Ireland
Shields, Kevin	CTO & Co-Founder, Crafting Tomorrow	Spain
Singh, Arnika	Director of Policy, Programs, and Research, Social & Media Matters	India
Tenus, Samantha	Founder and CEO, DigiPalz	Canada
Tiwari, Sonia	Children's Media Researcher, Oki Pie	United States
Tristan, Krystal	Founder, OneHaven	United States
Van Nood, Cristina	Chief Clinical Officer, ifeel online	Spain
Venansious, Steaphen Antony	Chief of Technology and Co-Founder, ChildSafe.dev	India
Wycoff, Tiffany	Founder, Generation Remix	United States

INDEX: People

INDEX: Topics

Citations [1] - [18]

[1] Reported by the Lancet Digital Health and multiple media outlets, noting the bill was introduced on November 21 and passed on November 29, 2024, with 'scarce opportunity for public consultation.' Source: The Lancet Digital Health, April 2025. URL: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(25)00024-X/fulltext

[2] Online Safety Amendment (Social Media Minimum Age) Act 2024, passed 29 November 2024, effective 10 December 2025. Source: Australian Parliament / eSafety Commissioner. URL:https://www.esafety.gov.au/about-us/industry-regulation/social-media-age-restrictions

[3] On September 4, 2025, Nepal's Telecommunications Authority ordered the blocking of 26 social media platforms including Facebook, YouTube, Instagram, and WhatsApp, after platforms failed to register under the Ministry of Communication and Information Technology. Source: Kathmandu Post, September 5, 2025. URL: https://kathmandupost.com/money/2025/09/05/confusion-as-nepal-bans-unregistered-social-media-sites. Also confirmed by Britannica, Al Jazeera, and Wikipedia.

[4] ChildSafeNet Nepal (2025), Enhancing the Role of Parents to Ensure the Safety of Children from Online Sexual Exploitation and Abuse in Nepal (mixed-methods study, ~900 parents, Kathmandu & Lalitpur); UNICEF Nepal & ChildSafeNet (2024), Nepal's Digital Generation. URL: https://ictframe.com/nepali-parents-never-discussed-online/

[5] World Economic Forum (2025), The Future of Jobs Report 2025 (global employer survey of 1,000+ leading companies representing over 14 million workers across 55 economies, examining employer expectations on AI, automation, skills demand, and workforce transformation to 2030). URL: https://www.weforum.org/publications/the-future-of-jobs-report-2025/

[6] This is the speaker's own research finding presented at the conference. For supporting literature, see: Wolak, J., Finkelhor, D., Mitchell, K. J., & Ybarra, M. (2008), Online "Predators" and Their Victims: Myths, Realities, and Implications for Prevention and Treatment, American Psychologist, 63(2), 111–128. URL: https://www.apa.org/pubs/journals/releases/amp-632111.pdf

[7] This is the speaker's own research finding presented at the conference. For supporting literature, see: Livingstone, S., & Helsper, E. (2010), "Balancing opportunities and risks in teenagers' use of the internet: The role of online skills and internet self-efficacy," New Media & Society, 12(2), 309–329. Ito, M., et al. (2009), Hanging Out, Messing Around, and Geeking Out: Kids Living and learning with New Media, MIT Press.

[8] This learning is grounded in practitioner led and community embedded research. While no single comparative benchmark study by Mindful Clicks Africa has been published, peer reviewed and institutional research provides strong supporting context. For supporting literature, see: Bhui, K., Basu, D., Nagpal, S., Mutiso, V., Pillai, R., Hadfield, K., Lauwrens, Z., & Ndetei, D. (2024), "Acceptability and feasibility of a brief intervention to enhance resilience among young people and their families in India and Kenya," Cambridge Prisms: Global Mental Health. UNDP Resilience Hub for Africa (2024), Annual Report 2024: A New Dawn for Integrated Resilience Building in Africa.

[9] This is the speaker's own research finding presented at the conference, developed through the CAC Framework (Creativity, AI, and Children), which examines how generative AI design patterns shape children's creative agency. For supporting academic context, see: Resnick, M. (2007), "Sowing the Seeds for a More Creative Society," MIT Media Lab. Cite Dr. Sonia Tiwari / Oki Pie.

Citations [10] - [19]

[10] This structural observation is documented in: Livingstone et al. (2017) 'Maximizing opportunities and minimizing risks for children online,' LSE/EU Kids Online; OECD (2021) 'Empowering Children in the Digital Age.' Cite Kevin Shields / Crafting Tomorrow.

[11] Supported by behavioral economics literature. See: Kahneman (2011) 'Thinking, Fast and Slow'; FCA (UK Financial Conduct Authority) research on scam vulnerability (2022); Australian Competition and Consumer Commission Scamwatch annual reports. Cite Aapti Institute / Mousmi Panda.

[12] This is the speaker's own research finding presented at the conference. The finding aligns with established radicalisation research showing that pathways into extremist communities are typically relational and belonging led before becoming ideological. For supporting literature, see: Moonshot CVE (2021), Radicalisation: The Role of the Internet, Moonshot Impact Report 2021. Berger, J. M. (2018), Extremism, MIT Press (Essential Knowledge Series). Kruglanski, A. W., Bélanger, J. J., & Gunaratna, R. (2019), The Three Pillars of Radicalization: Needs, Narratives, and Networks, Oxford University Press. Cite Alexandria Onuoha (speaker's academic research).

[13] ChatGPT launched on November 30, 2022 (OpenAI). URL: https://openai.com/blog/chatgpt Widely reported; no specific 'classroom guidance' existed at launch.

[14] Italy's data protection authority (Garante) temporarily banned ChatGPT in March–April 2023. Source: Garante per la protezione dei dati personali, March 31, 2023. URL: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832

[15] Google announced Gemini integration in Google Workspace for Education in 2024-2025. Reporting: EdTech Magazine, 2025; TechCrunch coverage of Google AI in Education 2025. Specific 'without prior notice' characterization is the speaker's own account. Cite Rocío Ribelles Zorita / IST.

[16] EU AI Act (Regulation (EU) 2024/1689), Annex III, Point 3: 'Education and vocational training' is classified as high-risk AI. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[17] Supported by: UNESCO (2023), Guidance for Generative AI in Education and Research (documents lack of pre-deployment safeguards and pedagogical vetting for AI tools in schools). OECD (2021), Digital Education Outlook (notes that digital tools often enter schools via procurement, not curriculum governance). National textbook adoption frameworks (e.g., the U.S., EU member states) consistently require multi-year review and approval cycles for instructional content. Cite Rocío Ribelles Zorita / UNESCO / OECD.

[18] Internet Watch Foundation (IWF) reported a 360% increase in self-generated sexual imagery of children aged 7–10 between 2020 and 2022. URL: https://www.weprotect.org/wp-content/uploads/Global-Threat-Asssesment-2023-Press-Release.pdf

[19] This learning is derived from the speaker's own applied research and testing conducted through OneHaven, examining the real-world effectiveness of commonly used parental safety and monitoring tools. The findings suggest that tools widely perceived by parents as protective often fail to function as expected in practice. Cite Krystal Tristan / OneHaven Research.

Citations [20] - [30]

[20] ECPAT, INTERPOL, and UNICEF Innocenti (2021). 'Disrupting Harm in Kenya: Evidence on online child sexual exploitation and abuse.' The report found that a substantial proportion of children did not know where to report or get help. The 61% figure is cited specifically in the conference document. The Kenya report documents high rates of non-disclosure and limited knowledge of reporting pathways. URL: https://www.interpol.int/en/News-and-Events/News/2021/Ground-breaking-insights-into-the-risk-of-online-child-sexual-exploitation-and-abuse-in-Kenya Full report: https://safeonline.global/wp-content/uploads/2023/12/DH-Kenya-Report_Revised30Nov2022.pdf

[21]] ECPAT, INTERPOL, and UNICEF Innocenti (2021). 'Disrupting Harm in Kenya.' The report documents caregiver responses to disclosure that included device confiscation. URL: https://safeonline.global/wp-content/uploads/2023/12/DH-Kenya-Report_Revised30Nov2022.pdf

[22] Sheng is a recognized pidgin/creole of Nairobi, documented by linguists. See: Githiora (2002) 'Sheng: Peer language, Swahili dialect or emerging Creole?' Journal of African Cultural Studies. URL: https://www.jstor.org/stable/3181415

[23] WEIRD framework coined by: Henrich, Heine & Norenzayan (2010). 'The weirdest people in the world?' Behavioral and Brain Sciences, 33(2-3), 61-83. URL: https://doi.org/10.1017/S0140525X0999152X

[24] Speaker's own research finding; supported by emerging literature. See also: Johnson et al. (2022) 'Ghost in the Machine? Pathologizing AI-based Mental Health'. Cite Ghowash Irshad / Montclair State University.

[25] Blacksky's own platform statistics. Cite Dr. KaLyn Coghill / Blacksky Algorithms directly. URL: https://blacksky.app

[26] Direct quotation from speaker. Cite: Michael Geoffrey Asia, Data Labelers Association (DLA), SHIELD Conference 2026. Supporting literature: Perrigo (2023) 'Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic.' TIME Magazine. URL: https://time.com/6247678/openai-chatgpt-kenya-workers/

[27] Speaker's argument: experiential expertise framing. Supported by: Hart's Ladder of Participation (1992); Percy-Smith & Thomas (Eds.) (2010) 'A Handbook of Children and Young People's Participation' (Routledge). Cite Lola Fisher / Gen Z Aotearoa.

[28] Supported by: World Economic Forum Future of Jobs Report 2025 (estimates 85M jobs displaced by 2025, 97M created); McKinsey Global Institute (2023) 'Generative AI and the Future of Work in America'; Goldman Sachs (2023) report estimating 300M jobs affected globally. Cite speaker / panel.

[29] Nwachukwu, N.M. (January 2026). Dataset of 565+ non-consensual Grok image requests documented at Trinity College Dublin's ADAPT Centre / AI Accountability Lab. Reported by The Guardian (Jan 8, 2026): 'Hundreds of nonconsensual AI images being created by Grok on X.' URL: https://www.theguardian.com/technology/2026/jan/08/grok-x-nonconsensual-images. See also: The Conversation (March 2026), University Times Dublin (March 2026).

[30] AI Forensics analysis cited in The Guardian (Jan 8, 2026) reporting 'nearly 6,000 requests per hour' to nudify women. URL: https://www.theguardian.com/technology/2026/jan/08/grok-x-nonconsensual-images. Also cited at BizBrief.ie and SPLC Center, Jan-Feb 2026. The Guardian, Reuters, multiple outlets, October-November 2022.

Citations [31] - [42]

[31] AI Forensics analysis: 'Between Christmas and 6 January 2026, two percent of these requests generated images depicting minors.' Reported in The Quint, January 2026. URL: https://thequint.substack.com/p/how-people-exploited-ai-using-groks

[32] Nwachukwu's The Conversation article (March 2026) states this directly. Supporting reporting: Musk's acquisition of Twitter in 2022 included large-scale layoffs of trust and safety teams, widely reported. Source: The Guardian, Reuters, multiple outlets, October-November 2022.

[33] Stanford Center for Research on Foundation Models (CRFM) (2024), Acceptable Use Policies for Foundation Models; Foundation Model Transparency Index analysis. URL: https://crfm.stanford.edu/2024/04/08/aups.html. Cite Stanford CRFM / Nwachukwu contribution.

[34] Speaker's own forensic audit research. Supporting context: Fuerte's PAUSI framework and user transcript analysis. Cite Giselle Fuerte / Being Human with AI. See also: Harrington et al. (2023) on AI companion dependency.

[35] The broader tech sector's Brussels lobbying budget increased more than 55% since 2021, reaching €151 million annually. Source: Corporate Europe Observatory & LobbyControl (October 2025). 'Big Tech lobby budgets hit record levels.' URL: https://corporateeurope.org/en/2025/10/big-tech-lobby-budgets-hit-record-levels. Also reported by Euronews, October 29, 2025. Note: The 55.6% figure specifically may refer to since 2021 (pre-Act passage), not solely post-passage.

[36] Avaaz (2023), Human Rights in the AI Act (public petition and open letter to EU trilogue negotiators opposing voluntary-only Codes of Practice and advocating for Fundamental Rights Impact Assessments). URL: https://secure.avaaz.org/campaign/en/human_rights_in_the_ai_act/

[37] Kenya National Commission on Human Rights: at least 60 killed by October 31, 2024, and over 80 enforced disappearances documented by December 2024. Government figures (September 2024) cited 132 officially missing. Source: Amnesty International Kenya Report 2024; CIVICUS Monitor (2024); URL (Amnesty): https://www.amnesty.org/en/location/africa/east-africa-the-horn-and-great-lakes/kenya/report-kenya/. Human Rights Watch (November 2024). URL (HRW): https://www.hrw.org/news/2024/11/06/kenya-security-forces-abducted-killed-protesters

[38] Kenyan Court of Appeal (2026), Bloggers Association of Kenya v. Attorney General, Court of Appeal ruling declaring Sections 22 and 23 of the Computer Misuse and Cybercrimes Act unconstitutional due to their overbreadth and use in suppressing dissent, particularly following the 2024 Finance Bill protests. Primary source: Kenya Law Reports (kenyalaw.org). Supporting documentation: CIVICUS Monitor (2024); Human Rights Watch reporting on post protest enforcement.

[39] University of Houston Graduate College of Social Work. Cite Jason Fernandez / 60 Watts of Clarity. URL: https://www.uh.edu/socialwork/

[40] Direct quotation from speaker. Cite: Jason Fernandez, SHIELD Conference 2026.

[41] Cite Sara Portell / HCRAI, Portugal. Framework documentation available from HCRAI directly.

[42] UK Department for Education. 'Generative AI: safety guidance for education providers.' January 2026. URL: https://www.gov.uk/government/publications/generative-ai-safety-guidance-for-education-providers

Citations [43] - [52]

[43] Problem Gambling Severity Index (PGSI) is the established tool this adapts. Source: Ferris & Wynne (2001). 'The Canadian Problem Gambling Index.' Canadian Centre on Substance Abuse. Cite Giselle Fuerte / Being Human with AI for the AI adaptation.

[44] Speaker's own research. Supporting literature: Shing et al. (2018) 'Expert, crowdsourced, and machine classification of suicide risk from online posts' (ACL Anthology); Zirikly et al. (2019) CLPsych shared task on suicide risk assessment. Cite Saba Oji / HolisticMindAI.

[45] Speaker's own program research. Cite Samantha Tenus / DigiPalz Canada.

[46] Supported by: World Economic Forum (2024), Global Cybersecurity Outlook 2024. Assesses global cyber inequity and documents structural disparities in access to cybersecurity capacity, safety audit infrastructure, and assurance mechanisms, particularly affecting smaller organizations and actors in emerging markets. World Economic Forum (2025), Global Cybersecurity Outlook 2025. Extends analysis of cyber resilience gaps and systemic risk arising from unequal access to security, governance, and compliance infrastructure across regions and sectors.

[47] Polyvagal Theory (Stephen Porges, 2011) and somatic therapy literature support this. See: Porges, S.W. (2011). 'The Polyvagal Theory.' Norton & Company. Also: van der Kolk, B. (2014). 'The Body Keeps the Score.' Viking. Cite Claire Calfo.

[48] Supported by: Eisenberg et al. (2004) 'Correlates of Fruit and Vegetable Consumption in College Students,' and particularly: Fulkerson et al. (2006) 'Family dinner meal frequency and adolescent development,' Journal of Adolescent Health. The Protective Value of Family Meals (CASA Columbia, 2012).

[49] Supported by: Putnam, R. (2000). 'Bowling Alone: The Collapse and Revival of American Community.' Simon & Schuster. Cite Neil Milton / The Table Talk Project.

[50] Supported by: Livingstone, S. & Blum-Ross, A. (2020) 'Parenting for a Digital Future.' Oxford UP; and broader literature on the 'digital divide.' Cite Angeline Corvaglia / SHIELD – APS

[51] This finding is based on the speaker’s own pilot data using the PAUSI (Problematic AI Use Screening Instrument). The data derives from a self selected respondent sample and is presented with appropriate methodological caveats. Within this sample, approximately 70% of respondents scored within the problematic use range. Cite Giselle Fuerte / Being Human with AI.

[52] Speaker's own platform statistics. Cite Steaphen Antony Venansious / ChildSafe.dev. URL: https://childsafe.dev

Citation Notes & Methodology

This citation reference was compiled at the time of the conference document's publication. Claims fall into two categories:

1. Independently verifiable facts: cited with primary or reputable secondary sources.
2. Speaker's own research / frameworks: where no external citation exists, the speaker and their organization are cited as the primary source, with supporting literature provided where available.

Compiled: April 2026 | For questions, contact the SHIELD editorial team (contact@shieldthefuture.com).