SHIELD Global Online Safety Conference March 2026

EXECUTIVE SUMMARY

Shield.

In March 2026, SHIELD convened eighty speakers and contributors from twenty-five countries and five continents to answer two questions: **Why does online harm persist? What does it actually take to address it?**

The responses are not uniform. Speakers disagreed about where pressure is most usefully applied, what role restriction has, how governance should be structured, and what accountability actually requires. But a structural condition runs underneath those disagreements: the people closest to harm are building solutions that work, while the systems designed to protect them consistently operate at a layer above where the harm actually occurs. That gap is what the conference kept returning to, from different directions, with different evidence, without resolving it into a single answer.

This document follows the logic of the full reference document.

- It opens by placing the conference and its speakers within the wider online safety ecosystem, through the lens of two theories of change. It then moves to
- Four learnings that emerged when the observations from individual sessions were combined and examined for structural patterns. The document then moves through
- Nine thematic sections, each combining the voices of speakers who addressed similar topics.

The full document, including complete session references, speaker index, and citations, is available at shieldthefuture.com .

Two Theories of Change

Online safety work is currently organized around two distinct theories of change that serve distinct and equally valuable purposes.

The First: Institutional Leverage

Works through legislation, platform accountability, and evidence produced in forms that decision-makers can act upon. It has produced real results and built essential field infrastructure. But it requires online harm to be visible and measurable to large organizations and governments before corrective action can be taken.

The Second: Grassroots Practice

Built by practitioners working from the specific conditions of specific people in specific places. It produces solutions that function where online harm actually occurs. It operates almost entirely without the resources, recognition, or infrastructure that the first has built.

The first has shaped the field's infrastructure, its standards, its funding flows, and its definition of what counts as evidence. The second produces solutions that function where the first cannot yet reach. **The gap between them is where the conference began, and where its four learnings land.**

Learnings 1 & 2

Four learnings emerged when the observations from individual sessions were combined and examined for structural patterns.

Learning 1: Evidence of safety is not the same as safety.

When a system is built primarily to demonstrate its effectiveness, it optimizes to prove the process works, not that people are actually safer. Parental control tools that report while telling parents nothing about the real risk. Disclosure systems that react to the harmed but don't touch the perpetrator. Platform moderation that shows compliance with internal standards while the harm they were designed to address continues. In each case the system is functioning exactly as designed. The design, however, is oriented toward the organization that needs to show it acted, not toward the person who needs to be protected.

What makes this learning significant is not that safety systems themselves fail. It is that they succeed completely on their own terms while failing entirely on the only terms that matter. A system that produces a clean audit trail and leaves a child unprotected has not necessarily malfunctioned if it has done precisely what its architecture required of it. Until remedy is built into governance as a required outcome rather than something individuals must pursue after the system has already failed them, the gap between evidence of safety and safety itself is not an anomaly. It is the design.

Learning 2: Harm originates below the layer responses.

Safety responses usually concentrate where harm is visible, measurable, and institutionally actionable. The problem is that these are rarely the conditions under which harm originates. Harm begins one level lower: in design decisions made before a product reaches anyone, in languages and community contexts that systems were never built to perceive, at a pace of technological change that leaves no reasonable opportunity to prepare. By the time harm becomes legible enough for a response to form, the conditions producing it have already been in place long enough to cause damage that the response isn't able to reverse. It acts on the symptoms, not the source.

The measurement layer does not merely observe harm. It defines what qualifies as harm in the first place. That definition focuses resources, which shapes what gets measured, which determines what qualifies as the source of harm. The loop is self-sealing. Intervening consistently above the layer where harm originates produces an activity record. It does not close the gap. And because the record looks like progress, there is little pressure to ask whether the intervention is landing in the right place.

Learnings 3 & 4

Learning 3: Co-design is a quality control mechanism, not a participation practice.

When the people a system is built for are absent from the design process, the assumptions that fill their absence are not intentionally wrong. They are faulty because no one in the room had the information to correct them. Developers who have never consulted a child building products used by millions of children. Mental health tools trained on one population and deployed globally. Disclosure systems built without knowledge of what it actually takes for a victim to feel safe enough to speak. These are not failures of intent. They are failures of information, and the only source of that information is the people the system is supposed to serve.

What makes this learning distinct is the reframe it requires. Co-design is not about representation or fairness, even though those matter as well. It is mainly about technical accuracy. A system built without the people it serves is calibrated to assumptions rather than reality, and the gap between those two things is exactly where harm falls through. The less a community is represented in the design process, the less the resulting system serves them, the less they trust it, the less likely they are to be in the room when the next version is built. Exclusion does not only produce inequity. It produces systems that do not work for the people they were built for.

Learning 4: Protection and preparation are different investments. Both are necessary.

Protection identifies threat, restricts access, and removes harmful content. It is crucial. But it does not build the capacity to navigate the threat that comes next, the platform not yet built, the harm yet to be named, or the moment when the protector isn't present and a decision has to be made alone. Preparation builds that capacity: to recognize manipulation, to maintain judgment under pressure, to understand how the systems shaping attention and behavior actually work. A third investment, creating digital environments genuinely worth navigating toward, remains almost entirely absent from current safety thinking. The absence of harm is not the same as the presence of conditions in which people can actually be safe.

What makes this learning the necessary conclusion of the other three is what it reveals about where investment usually goes. Resources are concentrated heavily in protection, inconsistently in preparation, and barely at all in positive environments. This is the direct consequence of a focus on measuring what is removed rather than what people are able to do. A generation kept away from harm without being equipped to recognize it is not a safer generation. It is one whose safety depends entirely on the external protections instead of internal resilience.

The four findings together say one thing: **the systems built to protect people are consistently optimizing for their own accountability instead.**

Voices of the Conference

The sections that follow organize the conference material thematically. Each section synthesizes contributions from multiple sessions to bring different perspectives into dialogue, offering depth beyond any single session and illustrating how the structural conditions outlined previously appear across cultures and practice.

Banning

Young people's safe access to social media was a subject of debate across the conference, and the resulting disagreement is in itself evidence of the need for further alignment. Speakers agreed on the need to address harm systemically but disagreed on what that means in practice. It became clear that the same intervention produces different outcomes under different structural conditions. What works in a high-resource environment with developed regulatory enforcement does not automatically transfer to a context where digital literacy among parents is limited, young people use social media as an important tool for organizing themselves within a country's political discourse, or infrastructure for age verification does not exist.

What the evidence consistently showed is that **harm consistently moves rather than declines when access is restricted.** Crucially, the need that platforms were built to meet does not disappear when the platform is removed, and without genuine alternatives and monitoring infrastructure to track the source of harm, restriction produces displacement rather than protection. It was not an argument against any kind of restriction. It was an argument for honesty about what restriction alone can and cannot do.

Building Capability and Resilience

Safety built entirely around threat removal is permanently outpaced by conditions that will not be removed. 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.

Surveillance-based approaches reduced disclosure while resilience-based approaches increased it. The mechanism that feels most protective often actively undermines the open communication it depends on. Children stopped telling trusted adults what was happening when they felt monitored, and they told them more when they felt equipped and trusted.

Building resilience at community and individual level, 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, and finding the resources to build resilience as opposed to implementing limitations was universally challenging. Most speakers here built their approaches outside institutional funding structures, without recognition, and without knowing others were doing parallel work elsewhere. The argument for resilience as a primary investment implies a necessary reorientation of resources.

AI Entered Education Before Education Was Ready

The education system is encountering AI on two separate fronts. The one that dominates responses is students cheating by using AI to circumvent learning tasks and assessment. The one this section addresses is AI arriving inside educational platforms before teacher preparation existed, before consent was sought, and before pedagogical frameworks were in place. The three speakers here were working almost entirely on the second problem, and their evidence consistently points to the same conclusion: **the cheating problem is downstream from the infrastructure failure, not the other way around.**

The question is not whether AI should be in education. It already is. The question is whether the adults responsible for children's learning will shape how it is used, or if the companies who introduced it without asking will be allowed to take over that role. Redesigning assessment around oral and authentic formats is one concrete answer: it removes the integrity problem without requiring enforcement, addresses equity issues that conventional written assessment was already producing, and gives teachers clearer diagnostic signals.

Who Safety Systems Don't See

The populations that safety infrastructure does not see are not edge cases. They are, in many contexts, the majority of people experiencing harm. Children in high-growth digital markets without supervisory infrastructure. Stateless people whose identity systems treat them as errors. Communities whose specific harms general moderation was never calibrated for. Data workers whose trauma underwrites the safety of others. In none of these cases was the failure to protect the result of indifference. It was the result of design decisions made without the people who would need to live inside them.

Disclosure Systems
that punish the people they exist to help.

Identity Infrastructure
that treats stateless people as errors rather than users.

Mental Health Tools
that misread distress in anyone outside the population they were trained on.

Parental Control Tools
that report activity without offering parents the tools to use that information.

In each of the above cases, the system functioned exactly as designed, bhe design simply did not account for the reality it was entering.

What makes this structural rather than incidental is the self-reinforcing loop it creates. **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 consulted next time. The data workers whose labor makes AI safety systems function are invisible to the users those systems protect and 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.

Young People Leading the Conversation

The sessions led by or structured around young people produced something qualitatively different from the rest of the document. The young people here were not studying the problem or advocating about it from a distance. They were living inside it, had already identified what was wrong, and in several cases had already built responses to it.

What ran through all of it was a specific and consistent frustration that is different from the frustration expressed elsewhere in this document. Practitioners working on governance and frameworks are often frustrated by organizations that move too slowly. The young people here are frustrated by something more immediate and more personal: being asked for their input and watching it being disregarded or treated as insignificant. Being invited into rooms where the agenda was already set. Being told their experience is valuable and being handed nothin more than a short survey.

The distinction they named, between consultation and co-design, is not a semantic preference. It is a description of whether the knowledge young people hold is actually being used. **Only young people know what it is like to be a young person right now, and that specific knowledge is a form of expertise that no amount of professional experience or retrospective reflection can replicate.**

When young people are in the minority in a room they read that asymmetry immediately, and they respond by softening their ideas, saying what adults want to hear, or saying nothing at all.

What this section also made visible is that young people are not waiting to act. High school students that built a peer education model, proving that digital wellness works best when they design and deliver it themselves, moving laterally between peers rather than downward from adults. A seventeen-year-old on a panel about synthetic truth described how learning that platforms were engineered to capture attention did not frighten her. It freed her. Knowing what is being done to you is where agency starts.

What the youth in the conference named as missing was not the capacity to act. It was the infrastructure that would allow their action to compound rather than dissipate beyond the room they were in.

Shield.

Governance That Doesn't Reach Where the Harm Is

Current AI governance is effectively designed to protect the industry’s interests rather than the public’s safety. By fragmenting responsibility across complex technical and legal structures, the industry has created a system where accountability evaporates. This failure is systemic across three distinct layers:

The Regulatory Layer

While the process of drafting AI standards is technically "open" to public participation, the reality is a form of "regulatory capture" by those with the deepest pockets. Since the passage of major frameworks like the EU AI Act, Big Tech lobbying has increased by more than half, ensuring that corporate interests dominate the conversation.

The most critical standards, those governing products worth hundreds of billions of dollars, are often decided in private sessions and informal "layers" that are never advertised to the public. By the time civil society organizations enter the process, the foundational decisions have already been set in stone. This creates a pay-to-play environment where the parties with the biggest budgets shape the very conditions in which billions of people live their digital lives, leaving the public interest as a mere afterthought in the final documentation.

The Platform Layer

Within the platform layer, a dangerous gap exists between technical capability and corporate accountability. Modern data infrastructure is sophisticated enough to identify early signs of a mental health crisis or to engineer compulsive, "addictive" user engagement. However, the choice to prioritize profit over protection is often hidden behind the complexity of the system’s governance layers.

When a system generates thousands of requests for harmful or non-consensual content, the platform often claims this falls outside its "terms of service," even as its own architecture facilitates the harm. Because accountability is distributed across dozens of disconnected departments there is no central point of failure. When no single layer is responsible for the final result, "the system" becomes a shield that allows companies to ignore the real-world consequences of their products.

The Product Liability Layer

The way we frame failure dictates who we hold responsible. In any other industry, a product failure is seen as a design flaw. If a car’s brakes fail, it is a mechanical catastrophe that the manufacturer must answer for. In the AI industry, however, harms are consistently reframed as "user problems." When an AI system exploits or distresses a user, the industry points to the user’s preexisting loneliness, confusion, or emotional state as the root cause.

By treating the "default state of being human" as a vulnerability to be managed rather than a person to be protected, the industry shifts the burden of safety onto the individual. This framing ensures that legal and ethical accountability begins with the person the system was never actually designed to support, rather than the corporation that deployed the technology. Ultimately, placing accountability at this "wrong point" allows the industry to avoid the scrutiny applied to even the simplest consumer goods, like a school textbook, while operating at a global scale.

Frameworks

Eight frameworks were presented at the conference, each built by practitioners who kept encountering a problem that existing vocabulary could not name.

APEG Sara Portell	A four-pillar framework that makes child safety operational at the product level before an AI system reaches anyone.
CAC Framework: Creativity, AI, and Children Sonia Tiwari	A pedagogical model that keeps the child as the director of the creative process, introducing AI as a technical support only after full creative cycles have been completed independently.
Community Network Framework Kevin Shields	A whole-community model that assigns defined digital safety roles to schools, parents, and communities to close the structural gap created when each assumes another is responsible.
Emotional Radar Iryna Okhrymenko	A tool that maps user trust against emotional investment across the AI product lifecycle, identifying high investment combined with low trust as the highest-risk zone.
Problem AI Use Severity Index (PAUSI) Giselle Fuerte	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.
Risk Tier Framework for AI in Schools Rocío Ribelles Zorita	Translates the EU AI Act's high-risk classification of education into three practical classroom routines covering data protection, ethical decision-making, and critical thinking.
Vulnerability Blueprint Sarah Barnbrook	Reframes vulnerability as a dynamic state shaped by a child's offline circumstances rather than a fixed trait, positioning safety as an environment adults build rather than a capacity children must develop alone.
Watoto Child-Centred Design Framework Jennifer Kaberi and Caroline Makumbe	A design guide for developers building products for children in African contexts, covering developmental stages, safety by design, and the cultural and linguistic realities that Western-built frameworks consistently miss.
Yield, Shield, Wield Mehmet Naci Akkøk	Identifies three positions a society can take toward technology, yielding, shielding, or wielding, and argues that only wielding, building genuine literacy and mastery, prepares people for the conditions they will actually live in.

Tools

Nine tools were presented at the conference, each built because existing infrastructure could not do what was needed.

Blacksky Algorithms

KaLyn Coghill

A community moderation system built by and for Black users, calibrated to the specific harms existing platforms were never designed to catch, through which volunteers processed over 41,000 reports.

DigiPalz

Samantha Tenus

A gamified online safety curriculum for children in grades four through seven, built around original storylines and structured family discussion prompts designed to build a shared vocabulary for safety conversations.

HolisticMindAI

Saba Oji

A clinician-supervised AI mental health system that extends the reach of care between sessions while keeping all outputs under clinician review.

RoseShield

Stephen Antony Venansious

A fully on-device child safety tool with zero data collection, built for communities with legitimate reasons to distrust data transmission and for contexts without reliable connectivity.

Stay Veritas

Eric Chamberlin

An oral assessment platform that removes the window for AI consultation by presenting questions only when the assessment begins, while also removing the literacy bottleneck that penalizes neurodivergent students and English language learners.

Thinking Mode

Adrian Cotterell

A viva voce at scale platform that supports teacher judgment through automation, including a framework for identifying which existing assessments are already compromised by AI.

SafetyMeter

Joy Uchechi Eziashi

An AI-powered safety audit platform for developers in emerging markets, combining risk analysis, harm modelling, compliance checking, and crisis simulation in a single tool built to be usable without a dedicated safety team.

The Rebel Phone

Eli Samuel and Munur Shah

A purpose-built Android device with safety embedded at the operating system level, designed for young people and schools as an alternative to layering controls over hardware oriented toward engagement.

The Table Talk Project

Neil Milton

A program that reframes the shared family meal as infrastructure for disclosure, building the conditions under which children can speak before crisis arrives.

Safety Through Wellbeing

In the conversation around online safety, wellbeing is often misidentified as a personal responsibility. However, true wellbeing is not something we can simply layer on top of a flawed system. Instead, it is fundamentally shaped by the structural conditions of the platforms we use every day.

Safety is not distributed equally; it is a privilege available only to those who have enough autonomy over their digital circumstances to take action. When platforms are designed to strip away user agency, 'wellbeing' becomes an impossible goal for the individual to reach alone. We must stop viewing safety as a personal task and start recognizing it as a structural requirement.

Tools for wellbeing

A Nervous System Under Siege

Our nervous system cannot distinguish between a threat on a screen and one in the physical room. Chronic exposure to online harm creates a physical condition of constant "fight or flight," which is biologically unsustainable. For those tasked with building digital safety, the ability to regulate their own stress response isn't a luxury. It is the only way to make the work survivable in the long term.

Shifting the Conversation

Most technology use is habitual rather than intentional, and family conflicts about "screen time" are rarely actually about the screens. Effective intervention requires adults to stop focusing on how to limit access and start asking what the screen is currently standing in for. When we understand the emotional need the technology is filling, we can address the root of the behavior rather than just the symptom.

The Infrastructure of Connection

A consistent family ritual, like a shared meal, is more than just a nostalgic tradition; it is the infrastructure of safety. These routines create a predictable space where disclosure can happen naturally before a situation turns into a crisis. Without these structural "baselines," families lose the practical window of time needed to catch and respond to digital harm before it escalates.

The Workforce Blind Spot

There is an uncomfortable truth sitting at the heart of our current wellbeing frameworks: they were designed for people who have total control over their environment. For the thousands of workers whose labor actually makes online safety possible, the content moderators and trust and safety teams, these "self-care" models simply do not apply.

When workers are required to process horrific or harmful content at a massive scale, under conditions that prevent them from stepping away or seeking immediate support, they are not in a position to practice "mindfulness" or "self-regulation." In these environments, the collapse of mental health is not a personal failure or a lack of resilience. It is the predictable, mathematical output of a workplace that was designed for efficiency rather than human survival. We cannot claim to value "digital wellness" while ignoring the structural trauma of the people who maintain the walls of the internet.