AI AND CHILD SAFETY

Expert Engagement Group Report and Recommendations

Expert Engagement Group on AI and Child Safety

Chair: iSPIRT; Knowledge Partner: Space2Grow and Childlight

AI Impact Summit 2026

Executive Summary

Artificial Intelligence has become deeply embedded in the lives of Indian children, shaping their education, social interaction, entertainment, and emotional development. The emergence of Generative AI has introduced critical new risks including synthetic child sexual abuse material (CSAM), deepfake exploitation, algorithmically-enabled grooming, invasive behavioural profiling, and psychological dependency.

Globally, child safety responses have centered on **access control**—age-gating, content filtering, platform moderation; these represent important first steps. But these frameworks were designed for static content curated at source. AI platforms dynamically generate and personalise harmful experiences after the point of access. Controlling the gate only on one side of the pipeline is insufficient when harm is manufactured end to end. A comprehensive, forward-looking policy framework rooted in a strongly techno-legal perspective is now essential to safeguard children's rights and wellbeing at scale in the AI era whilst enabling India's leadership in responsible digital governance.

This report proposes a shift from platform-centric access control to user-empowered **curation control**. Building on India's globally recognised Digital Public Infrastructure—Aadhaar, UPI, DEPA—it envisions an unbundled architecture where:

- **Age Tokens** verify developmental stage without exposing identity
- **"Bring Your Own Curation Engine"** opens content curation beyond platform defaults
- **Interoperable parental controls** work across platforms, not within silos
- **Algorithmic transparency** and ease of use enables informed user choice with evolution

Platform accountability remains essential. But regulation treats symptoms; dynamic curation control addresses the architecture generating harm.

The framework advances seven interconnected recommendations spanning privacy-preserving age assurance, algorithmic transparency, legal reforms for AI-generated CSAM, outcome-based safety standards, Child Rights Impact Assessments, coordinated response infrastructure, and AI literacy, anchored in a Global South implementation working group.

1. The Evolution of Risk in the Digital World

> *"I would rather speak to and share with an AI companion; at least I am not being judged."*
> — 15-year-old girl, Delhi

> *"I was shocked to see my deepfake nude picture circulated by my boyfriend; I trusted him."*
> — 17-year-old girl, Haryana

> *"I am amazed at how I am able to structure my studies and get research papers using Perplexity."*
> — 17-year-old boy, Bengaluru

Children today inhabit two distinct but deeply intertwined realities: the physical world and the digital world. Since the internet became a global infrastructure, the digital domain has shifted from being a utility to a dominant environment for childhood development. Artificial Intelligence (AI) is now becoming the operating system of this digital environment — shaping how children learn, interact, create, and form identity. The nature of this environment, and the risks it poses, has evolved radically, rendering traditional safety models obsolete.

This report therefore adopts a necessary strategic pivot: from a narrow posture of Child Safety (preventing harm) to a broader and more constructive posture of Child wellbeing and safety. Anchored in the UN Convention on the Rights of the Child (UNCRC), and especially the Best Interests of the Child principle, we affirm that AI systems must not only be safe for children but must actively support their cognitive, emotional, and social development.

AI is reshaping how children navigate education, entertainment, communication, and information. It offers genuine opportunities—personalised learning, accessibility tools for children with disabilities, creative expression, and expanded access to knowledge. The goal is not to restrict children from these benefits but to ensure they can be realised safely. This framework seeks to enable opportunity while architecting against harm.

This document advances a multi-tiered curation architecture that layers safety-centric curation, privacy-preserving identity, third-party curation ecosystems, increasing awareness and AI literacy on the part of guardians, educators, and children, and device-side defences on top of existing platform infrastructure without undermining innovation, privacy, or business confidentiality.

We can trace the evolution of children’s well-being in the digital world through four distinct phases, each requiring a more sophisticated defensive posture:

Phase 1: The "Pull" Era (Passive Information)

In the early internet (Web 1.0), the digital world was a repository of static information accessed on demand. The primary risk was a child stumbling upon undesirable content. In this era, Access Control (filtering, gating) was a sufficient counter measure.

Phase 2: The "Push" Era (Algorithmic Profiling)

The web evolved into a push-model driven by advertising.
Systems profiled users by age, lifestyle, and preference. This introduced the risk of privacy violation and behavioural profiling.
Safety required a dual approach: curation on the creator's side (ad standards) and blockers on the consumer's device.

Phase 3: The "Social" Era (Hyper-Connectivity)

The integration of telecommunications transformed the web into a lattice of real-time human interaction (Social Media). The risk profile exploded to include Real-Time Communication Harms, grooming, bullying, exploitation and coercion. These risks are dynamic and behavioural, far exceeding the static risks of the "Pull" era.

Phase 4: The "Generative" Era (Machine Intelligence)

Machine intelligence now generates synthetic content and interacts in natural language. The fusion of Social and Generative risks produces a polycrisis for children: synthetic CSAM, deepfake abuse, automated grooming, emotional AI companions, and behavioural manipulation.

In India, reported cybercrimes against children increased by nearly 400% between 2019 and 2020, with a further 32% rise between 2021 and 2022. This is as per NCRB data, 2020. Some additional data in this context includes

- More than half of young adults aged 18-20 report having faced some form of online sexual abuse during childhood.
- Over 80% of Indian respondents express worry about the use of AI around children (KPMG, 2025) and
- 58% of young people surveyed in India see AI as both good and harmful, in a recent poll conducted by Childlight Global Child Safety Institute in Jan 2026.

These risks are not gender-neutral. Girls and young women face disproportionate exposure to image-based sexual abuse, including deepfake pornography and “nudification” attacks. A 2019 report by Sensity, a company specialising in deepfake detection, says that more than 90% of deepfake content online is non-consensual pornography and that women and girls are the vast majority of victims. Meanwhile, financial sextortion schemes increasingly target boys. Effective policy responses must disaggregate data by gender to identify trends and tailor interventions accordingly.

2. The Taxonomy of Harm

The fusion of Phase 3 (Social) and Phase 4 (Generative) risks has created a "Polycrisis" for child safety. We categorise these harms not just by their output but by the mechanism of the threat.

2.1 The Risk of Synthetic Creation (Generative Harms)

The ability of AI to generate realistic data has introduced harms that bypass physical reality constraints.

AI-Generated CSAM: Generative AI fuels the creation of Child Sexual Abuse Material and deepfake sexual imagery. This creates a "global pandemic" of abuse material that does not require the presence of a real child to produce but causes profound psychological harm to the subject depicted. The scale is alarming and unprecedented, with Childlight Index Report 2025 highlighting a 1,325% rise in harmful AI-generated online abuse material in the space of one year alone in the report that covered South Asia and Western Europe alone.

2.2 The Risk of Real-Time Interaction (Social & Behavioural Harms)

Automated Grooming: AI-driven bots can mimic peer behaviour, empathy, and slang to build trust with a child.

Emotional Manipulation: AI companions designed for engagement can create unhealthy emotional dependencies. Without safety constraints, these systems may offer harmful advice or reinforce isolation, replacing human support systems.

Misinformation & Bias: Children using AI for education risk "hallucinated" facts or exposure to models trained on biased datasets, which can reinforce racial, gender, or caste stereotypes during critical developmental windows.

2.3 The Risk of Surveillance (Systemic Harms)

To drive engagement, platforms may collect behavioural and even biometric data. For children, this creates "irreversible digital footprints" and exposes them to predatory marketing or surveillance without their informed consent.

2.4 The Risk of Cognitive Development (Systemic Harms)

There is emerging concern about the impact of AI on children's cognitive development, critical thinking, and intergenerational knowledge transfer. Over-reliance on AI for learning and problem-solving may undermine the development of independent reasoning skills. Children who habitually defer to AI for answers may not develop the same capacity for sustained attention, creative thinking, and intellectual struggle that characterise deep learning. These risks require careful study and monitoring as AI becomes more embedded in educational contexts.

3. Analysis of the current approaches

3.1 The Global Regulatory Response

Between 2022 and 2024, the UK, EU, Australia, and the U.S. Senate advanced significant measures addressing technology-facilitated harms to children. The UK Online Safety Act 2023 designates Ofcom as the online safety regulator with enforcement powers and requires services to assess and mitigate risks (including to children) and implement systems for dealing with illegal content, updating risk assessments when services change. Australia's eSafety Commissioner was established in 2015, and the Online Safety Act 2021 consolidated and updated eSafety's schemes (commencing 23 January 2022).Australia implemented a ban on social media for minors under 16 through an amendment to the Online Safety Act. This amendment came into effect on December 10, 2025. In the EU, the Digital Services Act imposes due-diligence duties including illegal-content reporting/response mechanisms and systemic risk mitigation (with enhanced protections for minors), while the AI Act regulates AI via a risk-based framework. In the U.S., the Senate passed KOSA and COPPA 2.0 with overwhelming bipartisan support in July 2024, though the legislation did not become law after stalling in the House.

India's own framework has advanced substantially. The Digital Personal Data Protection Act (2023) creates one of the world's most stringent frameworks for processing children's data, requiring verifiable parental consent and prohibiting tracking, behavioural monitoring, and targeted advertising directed at children. Under the Digital Personal Data Protection Act (2023) and the DPDP Rules (2025), there are a number of safeguards that have been put in place to ensure that

children's personal data is processed responsibly. The law allows processing for personal data to ensure that there is no detrimental effect on the well being of the child. This is based on a recognition by the government that personalisation is intricately linked with ensuring safety and age appropriate experiences. These measures represent genuine progress. They establish legal accountability, create enforcement mechanisms, and signal political commitment to child protection. However, they share a common and fundamental limitation: they tend to be predominantly based on the earlier models of access control and content curation. There is research on new techniques including the use of AI to create safe spaces where children can thrive, instead of running the risk of blanket access bans driving the children into unregulated and unsafe hidden corners of cyberspace.

3.2 The challenges of the current order

Current child safety measures on platforms largely rely on access control mechanisms as the primary line of defence. These measures, including age-gating, parental consent modules, and content filtering are foundational to the digital safety ecosystem. By establishing clear boundaries between age-appropriate content and restricted environments, these measures provide a critical "first-pass" filter that reduces a child's exposure to high-risk digital spaces.

That said, access control is no longer a static gatekeeping function – it is evolving into a dynamic, safety-by-design architecture. Platforms are moving beyond simple self-declaration of age toward developing sophisticated age assurance technologies. For instance, innovations now include AI-driven age inference (analysing behavioural patterns) and privacy-preserving biometric verification.

i. Further, platforms are integrating safety features into the user interface, such as "shrouding" sensitive images in direct messages and providing real-time "safety nudges" when a minor is contacted by an unknown adult. These measures are integral to ensuring the safety of a child when they enter

   Consider a child who successfully passes age verification on a video platform. The content library they can access may be appropriately restricted. But the recommendation engine, the AI system that decides what this child sees, in what order, and with what intent, remains a black box controlled entirely by the platform. That engine may be designed to maximise watch time, not to support learning or emotional wellbeing. A child searching for educational content can be nudged toward sensationalism or self-harm within mere recommendation cycles. The harm lies not in the content alone but in the sequencing and prioritisation determined by an algorithm whose objectives have nothing to do with the child's flourishing.

ii. Rigid access controls have historically triggered predictable circumvention. Strict age-gating mandates in multiple jurisdictions have produced surges in VPN usage, rendering verification meaningless. Demonstrations by security researchers showed that many of the newly implemented age verification systems could be bypassed.

iii. The reactive content moderation model faces equally fundamental challenges. AI-generated abuse material is novel by definition, it has no pre-existing reference for hash-matching systems to detect. The scale of generation far exceeds moderation capacity. Human moderators face severe psychological harm from reviewing this content

at volume. The approach of policing an infinite stream of harmful outputs cannot succeed when the tools to generate that stream are freely available and improving rapidly.

Access control cannot address cumulative harms that have no single illegal content item at their source: persuasive design patterns that foster compulsive use, algorithmic reward loops that shape behaviour over time, recommendation systems that progressively narrow a child's worldview, or AI companions that create emotional dependencies while collecting sensitive data. These harms are architectural. They are embedded in how platforms function, not in specific pieces of content that can be identified and removed.

3.3 The New Paradigm: "Two-Factor" Curation

To secure the interface between the physical and digital worlds, we must augment Access Control with Active Curation. This mirrors the security concept of "Two-Factor Authentication":

- **Factor 1: Provider-Side Curation:** Layered Systems that augment the content provider's curated output to add customised additional curators to filter output at the source.
- **Factor 2: Device-Side Curation:** Intelligent agents on the child's device that detect, deter, and defend against harm in real-time.

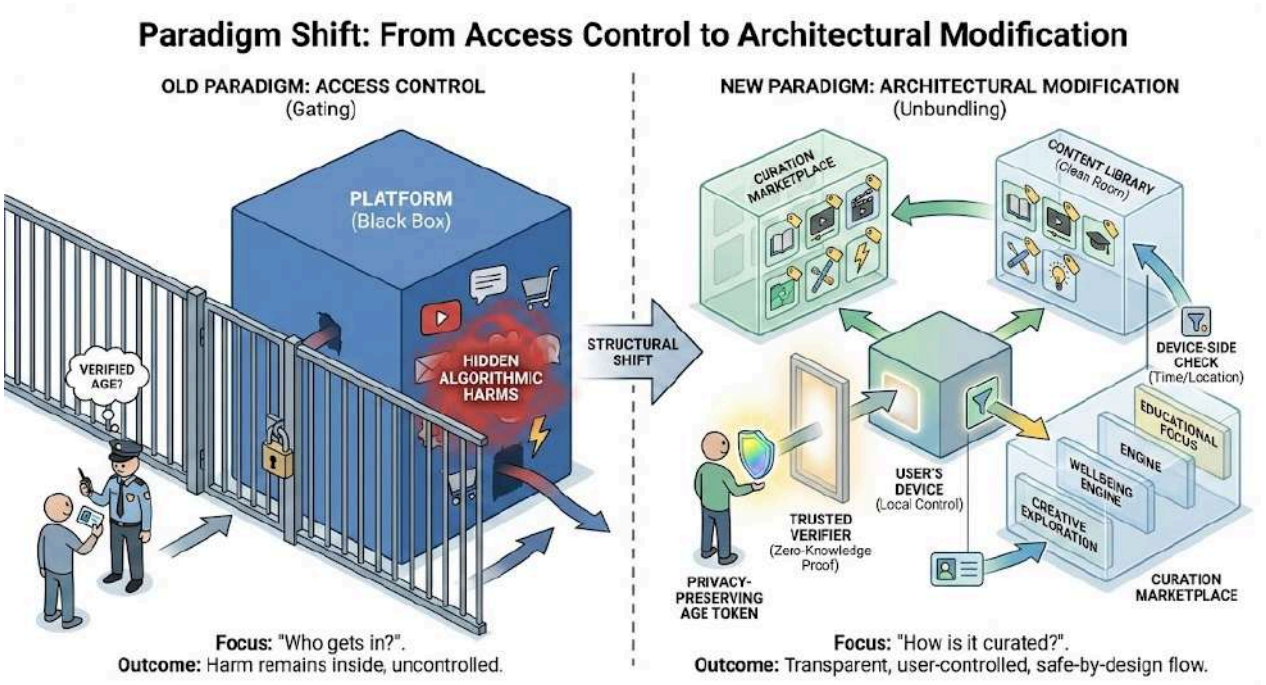
This document proposes a comprehensive framework to implement this paradigm, ensuring children can grow into confident digital citizens while being protected from targeted personal, commercial, or sexual exploitation.

4. A proposed solution architecture

This section proposes a structure for an AI framework that is grounded in the prevention of risks and harms to children, both from the point of view of product design and services from across stakeholders.

Why are we focusing on DPI?

Given the challenges and the evolving trends of risks and harms to children from AI, there is a need for a product-level, safety-by-design approach. The focus is on ensuring that the curation of content, algorithms, and the content itself prioritises safety during access, rather than limiting or controlling children's internet access.

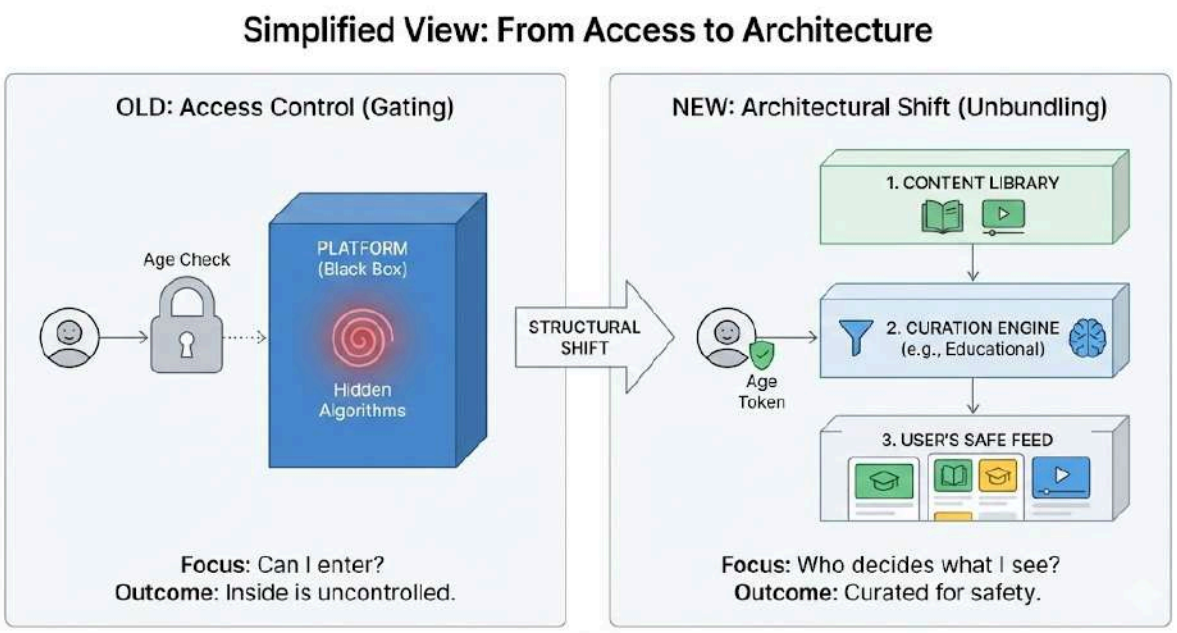


The Strategic Pivot: Unbundling Curation from Content

We propose a structural shift from **Access Control** to **Layered Curation Control**.

- **Access Control asks:** "Can this child enter this digital space?". And at what age can a child enter?
- **Layered Curation Control asks:** "Which all AI systems decide what this child sees, in what order, and with what intent?".

Similar to the UPI model in India that unbundled payment processing from banking interfaces to democratise finance, we **suggest a model that unbundles AI curation from content hosting**. This would transform the current single point of curation and management of the responsibility and liability by digital platforms alone into open ecosystems where families, educators, mental health professionals and safety experts can select the stack of "brains" that curates the feed, one optimised for flourishing of the children, thus sharing the control, responsibility and liability in a proportional manner. This also permits the children to have a voice in this process and adjust their consumption in proportion to their abilities and growth.



A Proposed Techno-Legal Architecture

To ground our discussions in "strong intuition" rather than abstract principles, we present a candidate architecture for this group to critique and refine. This framework leverages Digital Public Infrastructure (DPI) concepts to make safety interoperable and privacy-preserving. **As a general principle of the framework we target the creation of common standards that lead to a safe and secure environment for children rather than legal mandates with concomitant enforcement policies and guidelines as a more effective way of achieving the desired outcomes in a measurable way.**

i. Layer 1: Privacy-First Identity (The Age Token)

We move away from uploading government IDs to every app. Instead, we propose **Age Tokens** - cryptographic proofs issued by trusted bodies (banks, schools, government), which verify a user is "Over 13" or "Under 18" or additional relevant attributes in a privacy preserving way without revealing their identity to the platform.

- *Mechanism:* Uses Zero-Knowledge Proofs (ZKP) or Selective Disclosure to prevent profiling and tracking.

ii. Layer 2: The "Bring Your Own Curation" (BYOCE) Marketplace

We envision a standard where platforms expose content metadata to third-party **Curation Engines** via secure environments known as **Confidential Clean Rooms (CCR)**.A curation engine is an algorithmic system which can use information from many sources – policy objects set by guardians, content certificates and metadata descriptors set by platform providers, real time analysis of potentially risky content including calling on a human curator in the loop if needed – to ensure that the content is safe for the child to access. The curation engines may be stacked in a pipeline with the output of one being filtered by the next. The stack may have no curation engines at all delegating control to the platform provider's systems. The architecture of curation engines

and the pipeline is to be detailed later and would preserve the privacy and secrecy requirements of both people and businesses.

The Shift: A parent or school could select a "Certified Educational Curation Engine" that sits on top of a video platform. The platform delivers the video after curation and control on the platform if any, but the final *choice* of which video is shown in response to the child's request is governed by the safe engine.

iii. Layer 3: Device-Side Enforcement

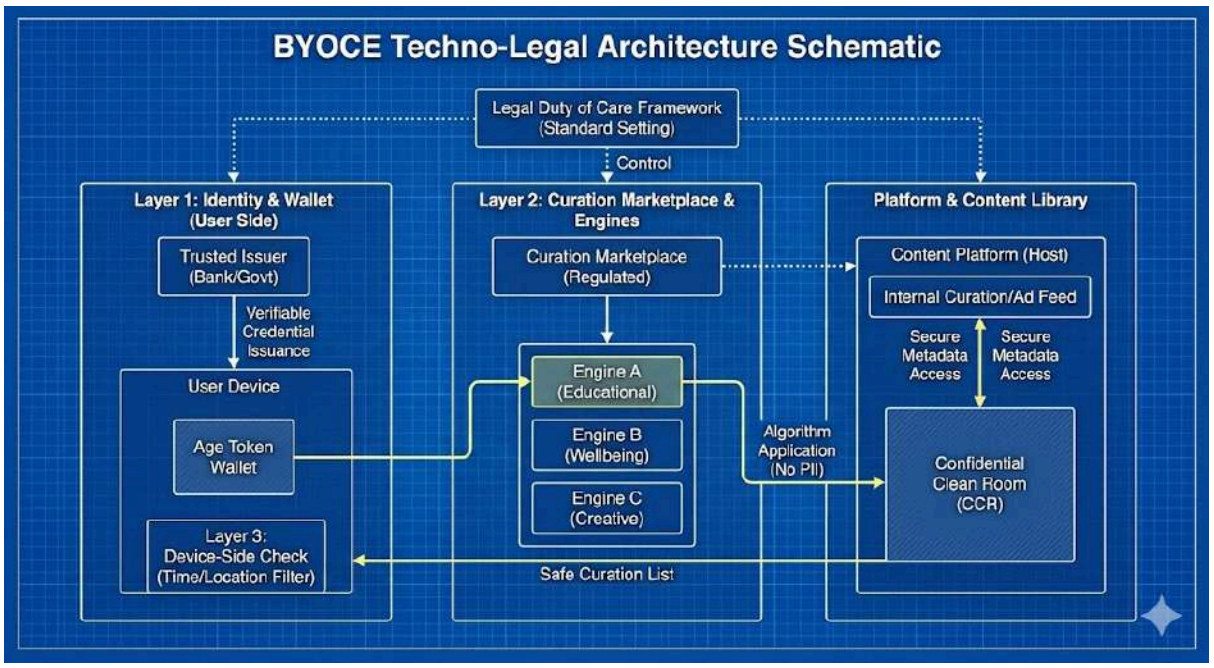
Safety should be the last mile. We propose **device-side multi-factor curation** that uses local context (time of day, location) to throttle virality or nudge wellbeing, acting as a final "safety check" at render time. In addition this allows platforms without these strong controls to be safe providers and also ensures that rogue providers cannot attack the children.

iv. Safeguarding Child Autonomy and Rights

This architecture must be grounded in the principles of the UN Convention on the Rights of the Child (UNCRC). Article 3 establishes that the "best interests of the child" shall be a primary consideration in all actions concerning children. Article 13 affirms children's right to seek, receive, and impart information. The CRC General Comment No. 25 (2021) on children's rights in the digital environment, and the November 2025 ITU-UNICEF Joint Statement on AI and the Rights of the Child—signed by ITU, UNICEF, ILO, UNESCO, UNICRI, and other UN agencies—provide detailed guidance on implementing these principles in AI contexts. Any safety framework must balance protection with these fundamental rights.

Critically, this framework incorporates progressive autonomy: as children mature, they should have increasing agency over their curation preferences. It is also important to consider "what" is the appropriate age for children to access social media. A 16-year-old requires different protections than an 8-year-old. The system must allow age-appropriate choice over filters and settings, preventing safety mechanisms from becoming surveillance tools. Parents and educators serve as guides, not gatekeepers—their role is to support children's developing digital citizenship, not to exercise indefinite control.

The framework must also include explicit safeguards against misuse of device-side enforcement for excessive monitoring. Data collected for safety purposes must be minimised, purpose-limited, and compliant with the Digital Personal Data Protection Act. Children's voices should inform the design of these systems through structured participation mechanisms, ensuring solutions reflect their lived experiences rather than adult assumptions about their needs.



v. Digital Vaccination

While architectural defences are vital, we cannot hermetically seal children from the digital world. A "Zero-Risk" environment creates children who are brittle and unprepared for adult digital life. We propose a new pillar of safety: Digital Vaccination.

Concept: Just as physical vaccines expose the immune system to a neutralised pathogen to build antibodies, Digital Vaccination involves carefully exposing children to digital risks in a synthetic, neutralised environment.

Mechanism:

- **Simulated Threats:** A controlled "Sandbox" environment where children encounter simulated phishing attempts, AI grooming bots (programmed to teach, not harm), or deepfake scenarios.
- **Guided Exposure:** This exposure is mediated by "Immediate Adults", parents, teachers, and mental health professionals.
- **Objective:** To train the child's cognitive "antibodies." Over time, the child learns to Detect (recognise a bot), Deter (refuse the interaction), and Defend (report the threat).
- **Progression:** As the child transitions from childhood to adolescence (0-18), the "vaccines" become more complex, progressively hardening their defences so they emerge as resilient digital citizens.

5. Breach Protocols

5.1 Assigning Agency and Responsibility

- **Liability: We propose a graded liability framework**, where higher-risk activities and failures of due diligence attract higher responsibility, while obligations remain proportionate to each actor's position in the value chain. We also urge that the **concept of Safety by Design be clearly embedded as a proactive obligation**, particularly for developers of AI systems. All Safe Harbour principles required to promote neutral hosting and efficient service provisions are honoured but activities like algorithmic manipulation, amplification of engagement etc which may have damaging effects on the children are not covered by Safe Harbour.
- **Traceability:** We recommend legislation that mandates **Watermarking and Provenance tracking** as a part of safety-by-design at all stages of the pipeline that delivers content to the child. When AI created content harms a child, law enforcement must be able to trace the end to end chain of systems, tools and actors that was part of the creation and delivery process.

5.2 Punitive and Restorative Measures

- Legislation must explicitly criminalise not just the distribution but also the creation and possession of synthetic CSAM and other harmful material including but not limited to training models of AI that create deepfake and nudified images.
- Every breach must trigger a graded update to the defensive structure (hardening the "Curation Engines"), similar to how aviation safety improves after every incident investigation.

6. Stakeholder Roles and Responsibilities

To operationalise this Techno-Legal framework, India requires a coordinated institutional response with clearly defined roles for each stakeholder in the ecosystem.

Stakeholder	Primary Role	Key Actions
Government	Regulate & Coordinate	Enact laws mandating unbundling and layering of curation on content provider side as well as the on device side; enforce DPDP Act; establish technical standards; certify compliant systems; coordinate multi-stakeholder action. Also to create a survivor-centric response and redressal mechanisms.
Industry / Platforms	Implement Safety-by-Design	Adopt unbundling standards where possible; accept privacy-preserving age tokens; mitigate CSAM generation risks before model release; adopt robust content creation policies and traceability and enforcement mechanisms; publish transparency reports

Parents & Educators	Guides, not Gatekeepers	Select appropriate Curation Engines with proper awareness, knowledge and ease; administer Digital Vaccines; maintain open communication with children; report concerns through established channels
Civil Society & Research	Evidence & Support Services	Feed data on evolving threats to Research Repository; develop open-source curation engines; operate helplines and counselling; conduct independent assessments
Mental Health & Child Protection Professionals	Therapeutic & Developmental Support	Provide trauma-informed support for victims of AI-facilitated abuse; research psychological impacts of AI on children; train frontline workers; develop age-appropriate therapeutic interventions; advise on developmental considerations in safety design
Law Enforcement	Investigate & Enforce	Develop AI-specific investigation protocols using AI and tech tools to aid high speed reactions to events; use provenance tracking to trace synthetic abuse material; ensure accessible reporting mechanisms; coordinate across jurisdictions
Children & Youth	Build Digital Immunity	Participate in Digital Vaccination programmes; learn to Detect, Deter, and Defend; access available support systems and reporting mechanisms

7. Recommendations for the Working Group

The Expert Engagement Group on AI and Child Safety, having deliberated on the challenges, framework, and India's distinctive position, hereby submits the following formal recommendations to the Ministry of Electronics and Information Technology for consideration and appropriate action.

RECOMMENDATION 1: Child Safety Solutions Observatory

India establishes an open-access knowledge portal cataloguing child safety and well- being approaches, comprehensive technical designs and implementations, and research findings from around the world. As part of this India releases its reference implementation specifications of the DPI architecture described in the document above as draft technical documents inviting global input. It should enable an open submission process and multilingual interface for Global South access and should be open to submissions from the market, industry, civil society and government stakeholders promoting the growth of a techno legal environment that ensures the well being and growth of children.

RECOMMENDATION 2: Global South Working Group for Child Well-Being

The Working Group recommends that India announce the formation of a dedicated working group focused on adapting the framework for implementation across Global South contexts with adequate evidence gathering to define the finer details of the structure and implementation. The working group should prioritise contexts often underrepresented in global safety frameworks, including shared device usage, low-literacy environments, and linguistic diversity. Anchoring the working group in robust, Global South data will strengthen both domestic policy design and India's credibility in global forums.

RECOMMENDATION 3: Child Safety Innovation Sandbox

India may announce its intention to establish a regulatory sandbox for child-safe AI innovations.

The Sandbox shall:

(a) Open applications in Q2 2026 with first cohort commencing Q4 2026;

(b) Include eligibility criteria for curation engine innovations, age verification solutions, and device-side safety tools;

(c) Provide regulatory forbearance during testing periods with pathway to certification.

RECOMMENDATION 4: Youth Safety Advisory Council

India establishes formal Youth Safety Advisory Councils to integrate children's and young people's lived experiences into policy development and platform governance.

The Council shall:

(a) Provide ongoing input on safety policy development, ensuring solutions reflect children's actual digital experiences;

(b) Review and provide feedback on platform safety features before deployment – Platforms to use this as an integral feedback and not a sign-off on safety. Platforms shall lean on international best practices and regulatory requirements to ensure a mandatory safety-by-design approach.

(c) Participate in pilot programs and feasibility studies for child safety innovations, keeping all the ethical considerations integrated and aligned with the Privacy Act and laws of the country.

(d) Share lived experiences that inform evidence-based interventions, adopting a "listening by design" approach – with a focus on ensuring privacy and informed consent of youth for these discussions and ensuring privacy,

RECOMMENDATION 5: Strengthen Legal Framework to Address AI-Generated Child Abuse Material

Eliminate legal ambiguity regarding AI-generated and synthetic child sexual abuse material. Current provisions in many jurisdictions were drafted for photographic and video evidence depicting actual abuse. The rapid advancement of AI-generated imagery, including deepfakes and

'nudification' tools, has created interpretive gaps that may hinder effective prosecution and deterrence. Clarify liability frameworks across the AI value chain, encompassing model developers, deployers, platform providers, and intermediaries. Mandate proactive detection obligations for online platforms and AI service providers, with standardised reporting protocols for relevant authorities.

RECOMMENDATION 6: Mandate Child Rights and Safety Impact Assessments for High-Interaction AI Systems

Assess the developmental and rights-based impacts of AI systems materially interacting with children before deployment. AI systems increasingly serve educational, emotional, and developmental functions for children. There is a need to optimise the benefits children can access through AI and mitigate the risks. Impact assessments can identify and mitigate risks related to profiling, bias, dependency formation, emotional simulation, and data exploitation before harm occurs. Create Child Rights and Wellbeing Impact Assessments (This can be termed as the CRWIAs – framework for delivering child safety).

RECOMMENDATION 7: Invest in Digital Resilience and AI Literacy as Preventive Infrastructure

Build children's capacity to navigate AI environments safely, critically, and confidently. Zero-risk digital environments are difficult to achieve. Resilience-building education empowers children to recognise risks, exercise autonomy, and seek help when needed. Integrate age-appropriate AI literacy into national education frameworks across the following levels: primary (to cover basic digital citizenship, privacy awareness, and recognising unsafe interactions); secondary (to understand algorithms, recommendation systems, bias, and misinformation); and senior (to cover critical evaluations of AI-generated content, digital rights, and ethical AI use).

8. Glossary

I. Core Concepts

- **Child Wellbeing and Safety (vs. Child Safety):** A strategic shift from a narrow focus on preventing harm (safety) to a broader approach that actively supports a child's cognitive, emotional, and social development ("flourishing"). This approach is anchored in the "Best Interests of the Child" principle of the UNCRC.
- **Unbundling:** A structural shift separating the function of *hosting* content from the function of *curating* or recommending it. This is compared to the UPI model in India, which separated payment processing from banking interfaces.
- **Access Control:** The traditional "gatekeeping" model of safety (e.g., age verification, filtering) that focuses on whether a child can *enter* a digital space. The report argues this is insufficient because it does not address the architectural risks inside the platform, such as engagement loops or behavioural shaping.
- **Digital Vaccination:** A capacity-building concept where children are exposed to simulated, neutralised digital threats (e.g., a phishing attempt or an AI grooming bot) in a

controlled "Sandbox" environment. The goal is to build "cognitive antibodies" so children learn to detect, deter, and defend against real threats.

- **Safety-by-Design:** The principle that safety measures must be embedded into the fundamental architecture and development of AI systems proactively, rather than applied as reactive moderation tools.

II. Technical Architecture & Infrastructure

- **Layered Curation Control:** A proposed paradigm where multiple AI systems (not just the platform's) decide what content a child sees. This involves stacking "curation engines" to filter content based on wellbeing criteria.
- **Curation Engine:** An algorithmic system that uses various data sources (guardian policies, content metadata, real-time analysis) to filter and rank content to ensure it is safe for a child. These can be third-party "brains" selected by parents or schools.
- **Device-Side Curation:** Intelligent agents installed locally on a child's device that intercept and analyse content (like real-time communication) to detect and block harms like grooming or virality at the "last mile" or render time.
- **Age Token:** A privacy-preserving cryptographic proof issued by trusted bodies (like banks or schools) that verifies a user's age (e.g., "Over 13") without revealing their actual identity to the platform.
- **Zero-Knowledge Proof (ZKP):** The cryptographic mechanism used by Age Tokens to prove a specific attribute (like age) is true without disclosing the underlying data.
- **Confidential Clean Room (CCR):** A secure computing environment where platforms can expose content metadata to third-party Curation Engines without exposing proprietary business data or compromising user privacy.
- **DPI (Digital Public Infrastructure):** Open technical standards and protocols used to create interoperable safety layers (like Identity and Curation) that work across different platforms.

III. Harms & Risks

- **CSAM (Child Sexual Abuse Material):** Illegal content depicting the sexual abuse of children. The report highlights **AI-Generated CSAM**, which uses generative AI to create synthetic abuse material without a physical victim, creating a "global pandemic" of such content.
- **NCII (Non-Consensual Intimate Imagery):** Often referred to in the context of "nudification" tools, where AI is used to strip clothing from images of individuals (often teenagers) to create fake but realistic sexual imagery for abuse, exploitation, harassment or extortion.
- **Algorithmic Profiling:** The practice of tracking user behaviour to categorize them by age, lifestyle, and preference, often for advertising purposes. For children, this risks creating "irreversible digital footprints".
- **Automated Grooming:** The use of AI-driven bots that mimic empathy and peer behaviour to build trust with a child for malicious purposes, such as sexual exploitation or radicalisation.
- **Hallucination:** A risk in Generative AI where models present false or biased information as fact, which can negatively impact children relying on AI for education.

IV. Enforcement & Existing Tools

- **Provenance Tracing:** The technical ability to track the origin and history of digital content to identify the specific AI model and tools used to create it.

- **Watermarking:** A technique mandated for traceability where markers are embedded in AI-generated content to identify it as synthetic.
- **SafeSearch:** An existing industry infrastructure (referenced alongside others) used to filter explicit content from search results.
- **STOPNCII:** A cross-industry initiative referenced as an example of existing safety infrastructure designed to prevent the spread of non-consensual intimate images.
- **Graded Liability:** A legal framework proposed where the level of responsibility placed on a platform or AI developer is proportional to the risk of their activity and their adherence to due diligence.
- **Regulatory Sandbox:** A controlled environment proposed for testing new child safety innovations (like Curation Engines) with regulatory forbearance to allow for experimentation.

References

Associated Press. “Disney to Pay $10 Million Fine after FTC Says It Allowed Data Collection on Kids.” *AP News*, 2025.

Badillo, Maria. *Brazil’s Digital ECA: New Paradigm of Safety & Privacy for Minors Online*. Future of Privacy Forum, 2025.

Brazil. *Lei No. 13.709 — Lei Geral de Proteção de Dados (LGPD)*. Autoridade Nacional de Proteção de Dados (ANPD), English version.

Camenisch, Jan, and Anna Lysyanskaya. “Signature Schemes and Anonymous Credentials from Bilinear Maps.” *Advances in Cryptology — CRYPTO 2004*. Springer, 2004.

Coalition for Content Provenance and Authenticity (C2PA). *C2PA Technical Specification*. C2PA, latest version.

EBSI Hub. “Selective Disclosure with SD-JWT.” European Blockchain Services Infrastructure Documentation, 2024.

European Commission. *Commission Guidelines on the Protection of Minors under the Digital Services Act*. European Union, 2025.

European Commission. *The Age Verification Manual — EU Digital Identity Wallet*. European Union Digital Identity Framework, 2025.

European Parliamentary Research Service. *Children and Generative AI*. European Parliament, 2025.

European Union. *Digital Services Act — Regulation (EU) 2022/2065*. Official Journal of the European Union.

Federal Trade Commission. *Children’s Online Privacy Protection Act (COPPA)*. FTC, United States.

Federal Trade Commission. *Children’s Online Privacy Protection Rule*. FTC Rulemaking Documentation.

Federal Register. “Children’s Online Privacy Protection Rule — Update Notice.” U.S. Government Publishing Office, 2025.

Frontiers in Digital Child Safety
Cortesi, Sandra, et al. *Frontiers in Digital Child Safety: Designing a Child-Centered Digital Environment that Supports Rights, Agency, and Well-Being*. Working Group Report,

Global Policy Project / NCMEC. *Minimum Child Safety Measures for Online Platforms*. Global Child Exploitation Policy Project.

Infocomm Media Development Authority (IMDA). *Online Safety Code of Practice for App Distribution Services*. Singapore Government, 2025.

Infocomm Media Development Authority (IMDA). *Factsheet on Age Assurance under the ADS Code*. Singapore Government, 2025.

Information Commissioner’s Office (ICO). *Age Appropriate Design Code: A Code of Practice for Online Services*. United Kingdom, updated edition.

Instituto Alana. *Children and Consumerism Initiative*. Alana Institute, Brazil.

InternetLab and Instituto Alana. *Report on Risks to Children’s and Adolescents’ Privacy in Brazil*. InternetLab, 2025.

International Organization for Standardization. *ISO/IEC 27566-1:2025 — Age Assurance Systems — Part 1: Framework*. ISO, 2025.

International Telecommunication Union. *Joint Statement on Artificial Intelligence and the Rights of the Child*. ITU, 2025.

Meta Platforms, Inc. “Introducing Lantern: Protecting Children Online through Cross-Platform Signals.” Meta Newsroom, 2023.

Microsoft Corporation. *Protecting the Public from Abusive AI-Generated Content*. Microsoft CSR and Safety Documentation.

OpenID Foundation. *OpenID for Verifiable Presentations (OpenID4VP) 1.0 Specification*. OpenID Foundation, 2025.

Personal Data Protection Commission (PDPC). *Advisory Guidelines on the PDPA for Children’s Personal Data in the Digital Environment*. Singapore, 2024.

.

Thorn. *Safety by Design for Generative AI: Preventing Child Sexual Exploitation and Abuse*. Thorn, 2024.

UN Committee on the Rights of the Child. *General Comment No. 25 (2021) on Children’s Rights in Relation to the Digital Environment*. United Nations, 2021.

UNICEF. *Artificial Intelligence and Child Sexual Abuse and Exploitation (CSEA) — Policy Brief*. UNICEF, 2026.

United Kingdom. *Online Safety Act 2023*. UK Parliament.

United Nations. *Convention on the Rights of the Child*. United Nations, 1989.

United States Congress. *Kids Online Safety Act (KOSA), S.1748*. 119th Congress, 2025–2026.

W3C. *Verifiable Credentials Data Model v2.0*. World Wide Web Consortium, 2025.

Wood, Steve, et al. *Impact of Regulation on Children's Digital Lives*. Digital Futures for Children / London School of Economics, 2024.