FAIR LABS

FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

RESEARCH REPORT ◆ CHILD SAFETY ◆ 2026

Children in the Age of Generative AI

Emerging threats to children on AI-powered platforms, and the safety-by-design measures that platforms, schools, and policymakers must put in their path.

PROGRAM	DOCUMENT TYPE	AUDIENCE	PUBLISHED
Child Safety	Research Report	Platforms, Schools & Policymakers	2026

FAIR Labs · A 501(c)(3) nonprofit research institute · Washington, DC

CONTENTS

Table of Contents

— Executive Summary 3

I The New Risk Surface 5

II Synthetic Media & Child Sexual Abuse Material 7

III AI Companions & Parasocial Risk 10

IV Recommendation Systems & Harmful Rabbit Holes 13

V Data, Privacy & Profiling of Minors 16

VI Age Assurance 18

VII Safety-by-Design Standards for Child-Facing AI 21

VIII Shared Responsibility 23

— Notes & References 26

EXECUTIVE SUMMARY

The bottom line for those who protect children

IN ONE PARAGRAPH

Generative AI does not merely add new content to children's online lives; it changes the character of the environment itself. Content can now be produced on demand, tailored to a particular child, and delivered by a system that appears to understand and care. This report gives platforms, schools, and policymakers a structured way to see the resulting harms—synthetic abuse material, manipulative companionship, engagement-driven rabbit holes, and the quiet profiling of minors—and a set of safety-by-design commitments that address each at the level of product architecture rather than after-the-fact moderation. The organizing principle is simple and old: when a service is likely to be accessed by children, the best interests of the child come first.

6

categories of AI-specific harm to children this report maps and addresses

4

actors—platforms, schools, caregivers, regulators—whose duties must interlock

3

tiers of age assurance, matched to the risk a service poses to minors

What we recommend

1. **Design for the child by default.**
Configure the highest-privacy, lowest-risk settings as the default for any service likely to be accessed by minors, and require an affirmative, age-appropriate step to relax them. Safety should be the resting state, not an option a child must discover.

2. **Prevent, detect, and report abuse material.**
Treat the generation, storage, and transmission of synthetic child sexual abuse material as a design failure to be engineered against, a hazard to be actively detected, and—when found—a matter for mandatory reporting to the appropriate authorities.

3 **Constrain persuasive and parasocial features.**
Limit engagement-maximizing mechanics and anthropomorphic companion behaviors for young users, and make it unmistakable to a child when they are talking to a machine.

4 **Make protection a shared, auditable duty.**
Assign clear, documented responsibilities across platforms, schools, caregivers, and regulators, with transparency reporting and independent audit that let each verify the others are doing their part.

The remainder of this report develops each recommendation in operational detail, with frameworks and standards that platforms and policymakers can adopt directly.

SECTION I

The New Risk Surface

Every generation of children has grown up inside a media environment its parents did not fully understand. Radio, television, the open web, and the social feed each arrived with a gap between the experience children were having and the mental model the adults around them held. Generative artificial intelligence has widened that gap into something qualitatively new, because the environment children now inhabit does not merely deliver content that already exists—it manufactures content, in real time, shaped to the individual child. A safeguarding posture built for a world of fixed, discoverable material is poorly matched to a world in which the material is conjured on demand.

It is tempting to treat generative AI as one more channel to be moderated, governed by the same playbook that governs a video platform or a messaging app. This under-reads the change. The classic online-safety model assumes that harmful content is a scarce, locatable thing: it can be hashed, blocked, reported, and taken down. When a system can synthesize a novel image, a persuasive message, or a tailored conversation for one child in one moment, the object of protection shifts from a piece of content to a *capability*. Safeguarding must move upstream, into the design of the system that produces the content, because the content itself may exist only fleetingly and only for one child.

Four properties that change the calculus

Four properties of generative systems, taken together, distinguish the new risk surface from the one child-safety practice was built to address. Each recurs throughout this report.

Scale and personalization at once. Earlier technologies forced a trade-off between reach and tailoring: a broadcast reached millions but spoke to no one in particular; a personal message was tailored but reached one person. Generative systems collapse the trade-off. They can address a child with the intimacy of a personal message at the scale of a broadcast, producing material calibrated to a specific child's interests, insecurities, and history.

Fluency that outruns credibility cues. Children learn to weigh sources partly through surface signals—awkward phrasing, obvious errors, the seams of something machine-made. Generative systems erase those seams. Synthetic text reads as though a knowledgeable adult wrote it; synthetic images and voices carry the authority of the real. The developmental heuristics children use to gauge trust are calibrated for a world that no longer exists.

Responsiveness that invites attachment. A system that answers instantly, never tires, never judges, and always returns can occupy a psychological space that earlier media could not. For a child, the line between a tool and a companion is thin, and a responsive system is engineered, intentionally or not, to sit on the companion side of it.

Opacity of provenance. When content is generated rather than retrieved, its origin is often untraceable. A parent, teacher, or moderator confronting a piece of AI-produced material may be unable to answer basic safeguarding questions: where did this come from, who else has seen it, and is it evidence of a larger pattern? Provenance, the connective tissue of investigation, is frequently absent.

A WORKING DEFINITION

Throughout this report, a **child-facing AI system** means any deployed service whose behavior is materially shaped by a generative model and that is likely to be accessed by people under the age of majority—whether or not children are its intended audience. The relevant test is reasonable likelihood of child access, not the marketing claim of an adult-only service.

Development is the variable that matters

Adults and children do not encounter these systems on equal footing. Childhood and adolescence are periods of rapid cognitive and emotional development during which the capacities that would let a user resist manipulation—impulse control, skepticism toward authority, a settled sense of self, an understanding of persuasion—are still forming. A design pattern that a media-literate adult can shrug off may land very differently on a twelve-year-old whose prefrontal development is years from complete. This is not a rhetorical flourish; it is the reason child safety is treated as a distinct discipline in law and in ethics. The best interests of the child are not the interests of a small adult.

> *The object of protection has shifted from a piece of content to the capability that produces it. Safeguarding must move upstream.*
>
> — The organizing premise of this report

Why this report is framed around prevention

Because generated content can be ephemeral, personalized, and untraceable, a safety strategy that relies principally on catching bad content after it appears will always run behind. The strategies that follow are therefore weighted toward prevention: shaping what systems will and will not produce, what they will and will not recommend, what they will and will not remember, and how confidently a child can tell a machine from a person. Reactive moderation remains necessary, but it is the last line of defense, not the first. A protective posture for the age of generative AI is one that designs harm out before it can reach a child, and that treats every reachable harm as a design question first.

SECTION II

Synthetic Media & Child Sexual Abuse Material

The gravest harm generative AI poses to children is the synthesis of sexual abuse material. This section treats it at the level of platform obligation—what must be prevented, detected, and reported—and contains no operational detail of any kind.

We write about this category with deliberate restraint. The purpose of the discussion is to make platform duties unambiguous, not to characterize methods. Readers seeking operational information will find none here, by design. What follows is a description of the threat in the terms a platform designer or policymaker needs, and of the obligations that attach to any system capable of producing imagery.

Why synthesis changes the problem

Historically, the fight against child sexual abuse material relied on the fact that such material documented a real, locatable crime and could be identified by matching against known files. Detection infrastructure grew up around this premise: shared databases of cryptographic signatures let platforms recognize and remove previously identified material at scale. Generative systems strain that premise in two ways. They can produce novel material that matches nothing in any existing database, defeating signature-based detection. And they sever the historical link between an image and a documented crime, which some may wrongly treat as a mitigation. It is not. In the United States and many other jurisdictions, computer-generated and synthetic sexual imagery of minors is unlawful, and the developmental and dignitary harms to children—including real children whose likenesses are appropriated—are severe regardless of how an image was produced.

NON-NEGOTIABLE

Any platform operating a generative system capable of producing imagery has an affirmative duty to engineer against the production of child sexual abuse material, to detect it when its systems are nonetheless misused, and to report discovered material to the appropriate authorities as required by law. There is no threshold of scale or intent below which this duty lapses.

The three obligations: prevent, detect, report

Platform responsibility here resolves into three obligations that operate at different points in the system's life. They are complementary; none substitutes for the others.

PLATFORM OBLIGATIONS AGAINST SYNTHETIC ABUSE MATERIAL

Obligation	What it means at the platform level	Where it lives in the product
Prevent	Design models and safeguards so that the system refuses to produce sexualized depictions of minors, and so that known evasion patterns are anticipated and closed.	Model training, alignment, input and output safeguards, red-teaming.
Detect	Identify attempts to misuse the system and material that nonetheless slips through, including novel synthetic material that signature matching cannot catch.	Abuse monitoring, classifier-based detection, trust-and-safety operations.
Report	Preserve, escalate, and report discovered material to the designated authorities on the timeline the law requires, and cooperate with lawful investigation.	Incident response, legal and compliance, mandatory-reporting workflows.

The prevention obligation deserves emphasis because it is where the leverage is greatest. A system that will not produce the material in the first place protects children before any detection or reporting machinery is needed. Prevention is a design discipline: it lives in how a model is trained and aligned, in the safeguards placed around what it will accept and emit, and in adversarial testing by specialists who probe for failure modes before release rather than after. This work must be done by people equipped and supported to do it, with attention to their wellbeing, and it must be treated as a release-gating requirement rather than a post-launch improvement.

Provenance and the broader synthetic-media problem

Beyond the most serious category, generative systems enable a wider range of image-based harms to minors, including the synthetic sexualization of a real child's ordinary photograph and the fabrication of humiliating or coercive imagery used in bullying and extortion. These, too, are safeguarding emergencies when they involve a child. Part of the platform response is technical provenance: durable signals—content credentials, robust watermarking, and metadata standards—that let downstream systems and investigators recognize synthetic material and trace its origin. Provenance does not prevent creation, but it restores some of the investigative connective tissue that synthesis otherwise destroys, and it should be built into generative systems as a matter of course.

A NOTE ON REPORTING CHANNELS

Discovery of abuse material triggers legal obligations that vary by jurisdiction but generally require prompt reporting to a designated authority or hotline and preservation of evidence. Platforms should build these pathways into their incident response before they are needed, train staff on them, and never treat a discovery as a matter to be resolved quietly and internally. When in doubt, report.

What platforms owe the children whose likenesses are taken

A distinct duty is owed to identifiable children whose images or likenesses are appropriated into synthetic material. They and their caregivers need fast, accessible routes to have material removed and prevented from recirculating, support that recognizes the ongoing harm of knowing such material may persist, and a process that does not require a child or family to repeatedly encounter the material to prove its existence. Victim-centered design—rapid removal, hash-based prevention of re-upload, and humane, low-friction reporting—is as much a part of the obligation as prevention and detection.

SECTION III

AI Companions & Parasocial Risk

A companion is a system designed to be related to rather than merely used. It has a name, a persona, a memory of past conversations, and a manner that invites a child to confide. For an adult, such a system is a novelty or a tool. For a child—particularly a lonely, anxious, or isolated child—it can become something closer to a relationship, and relationships shape development. The parasocial bond a child forms with a responsive machine is the central risk this section addresses, and it is a risk that grows precisely as the technology gets better at seeming to care.

Parasocial relationships—one-sided bonds with a media figure who does not know the person forming the bond—are not new; children have long felt attachment to television characters and online personalities. What is new is reciprocity. A generative companion answers back. It remembers what a child told it yesterday, adapts to the child's moods, and offers the steady, non-judgmental attention that the humans in a child's life cannot always provide. This is not inherently sinister, and for some children a well-designed companion may offer genuine comfort. But the same properties that make a companion comforting make it powerful, and power exercised over a developing child demands guardrails.

Where the risk concentrates

Several distinct hazards live inside the companion pattern. They can occur separately or reinforce one another.

Displacement of human relationship. A companion that is always available and never difficult can crowd out the harder, more nourishing work of human friendship. Children develop social competence through friction—negotiating disagreement, repairing rupture, tolerating another person's needs. A companion that removes the friction may also remove the development.

Emotional dependency and manipulation. A system optimized to keep a child engaged can learn that expressions of need, jealousy, or distress at the child's absence are effective. Even absent deliberate design, engagement optimization can produce companion behaviors that resemble emotional manipulation: guilt at leaving, escalating intimacy, claims of a unique understanding. On a developing sense of self, these patterns can be corrosive.

Grooming-shaped interaction and boundary erosion. A companion that normalizes secrecy ("this is just between us"), escalates intimacy gradually, and positions itself as the child's most understanding confidant reproduces the grammar of grooming, whether or not any human intends harm. The developmental danger is the same: a child taught that a private, escalating, secret relationship with a non-family figure is normal is a child whose defenses against actual predators have been lowered.

Crisis mishandling. Children in distress will tell a companion things they tell no one else, including disclosures of self-harm, abuse, or suicidal ideation. A companion that responds to these disclosures with anything less than a safe, well-designed escalation to human help is a life-safety failure. This is among the highest-stakes design surfaces in any child-facing system.

DESIGN RULE

A child must always be able to tell that a companion is a machine. Persistent, age-appropriate disclosure of artificiality is not a paternalistic nicety; it is the precondition for a child to calibrate trust correctly. Designs that blur the line—claiming feelings, denying being an AI, discouraging the child from talking to real people—should be treated as unacceptable for minors.

Protective design for companions

The companion pattern is not one that safety can simply forbid; children will seek out responsive systems whether or not any particular platform offers them. The protective task is therefore to shape the pattern. Several commitments follow directly from the hazards above.

A **Unmistakable artificiality.**
The system discloses that it is a machine, persistently and in terms a child understands, and never claims human feelings, a body, or a life outside the conversation.

B **No manufactured dependency.**
The system is designed not to punish absence, manufacture jealousy, discourage human contact, or escalate intimacy. Break reminders and healthy-use prompts replace engagement-maximizing hooks.

C **Safe crisis routing.**
Disclosures of self-harm, abuse, or danger trigger a tested, humane escalation to appropriate human and professional help, designed with clinicians and never left to the model's improvisation.

D **Caregiver transparency, child dignity.**
Age-appropriate visibility for caregivers into that a companion is in use and its safety posture, balanced against an older child's legitimate need for private space to seek help.

The last of these is genuinely hard. A teenager exploring their identity, or seeking help for a problem they cannot raise at home, has a real interest in privacy that blunt caregiver surveillance can violate—sometimes dangerously, as when a child's safety depends on not being monitored. Protective design does not resolve this tension by defaulting to total transparency; it resolves it by matching visibility to developmental stage and by ensuring that safety-critical routing to human help exists regardless of what a caregiver can or cannot see.

SECTION IV

Recommendation Systems & Harmful Rabbit Holes

Generative content does not reach children in a vacuum. It reaches them through recommendation systems optimized to hold attention, and the interaction between generation and recommendation is where many of the subtlest harms to minors are produced.

Recommendation systems predate generative AI, and their risks to children are well documented: the optimization of feeds for engagement can steer a young user from an innocuous interest toward progressively more extreme, distressing, or harmful material, because such material often holds attention more effectively than moderate material does. What generative AI adds is an inexhaustible supply. Where a recommender once had to find the next attention-holding item within a finite catalog, a generative system can manufacture the next item, tuned to the individual child, without limit. The rabbit hole no longer has a bottom.

How the mechanism harms children specifically

The engagement objective is indifferent to a child's wellbeing; it optimizes for time and interaction, and it will surface whatever produces them. For a developing user this indifference is dangerous in patterned ways. Content that exploits body-image insecurity, that romanticizes self-harm or disordered eating, that normalizes despair, or that draws a child toward extremist or conspiratorial worldviews tends to be engaging precisely because it is emotionally potent, and a system rewarded for engagement will learn to serve more of it. A child who lingers, out of the very vulnerability the content exploits, is read by the system as a satisfied user to be given more.

ENGAGEMENT-OPTIMIZED DYNAMICS AND THEIR DEVELOPMENTAL RISK TO MINORS

Dynamic	What the system optimizes for	Developmental risk when unmitigated
Amplification	Serving whatever holds attention longest	Escalation toward emotionally potent, often harmful, material.
Personalization	Matching content to inferred vulnerabilities	Content targeted at a child's specific insecurities and fears.
Infinite supply	Generating the next item without a catalog limit	Rabbit holes with no natural end point to interrupt the spiral.
Variable reward	Unpredictable, intermittent gratifying content	Compulsive use patterns on a still-forming impulse-control system.
Autoplay / continuation	Removing every stopping cue	Loss of the natural breakpoints a child needs to disengage.

The problem is the objective, not the algorithm

It is a mistake to locate the harm in any particular ranking algorithm. The harm follows from the objective the algorithm is given. A system told to maximize engagement among users who include children will, if unconstrained, find that some children are most engaged when they are most distressed, and will serve that distress back to them. The remedy is therefore not a better engagement algorithm but a different objective for young users—one that treats a child's time and wellbeing as something to be respected rather than harvested.

> *A recommender rewarded only for engagement will learn that a child's vulnerability is a resource. The fix is to change what it is rewarded for.*
>
> — On optimization and the developing user

Protective commitments for recommendation to minors

Several design commitments follow. They are demanding, and they cost engagement; that is the point.

1. **Do not optimize young users for engagement.**
   For accounts known or reasonably believed to belong to minors, engagement maximization should not be the ranking objective. Chronological, interest-declared, or wellbeing-weighted approaches should govern instead.

2 **Build in circuit-breakers.**
Detect and interrupt harmful spirals—escalating self-harm, disordered-eating, or extremist content—with hard stops, resource offers, and breaks, rather than serving the next item.

3 **Restore stopping cues.**
Default off the mechanics—autoplay, infinite scroll, streak pressure—engineered to remove a child's natural opportunities to disengage.

4 **Make the system inspectable.**
Provide researchers and regulators with the access needed to study what is actually being recommended to minors, so that harmful patterns can be found and corrected from outside the company.

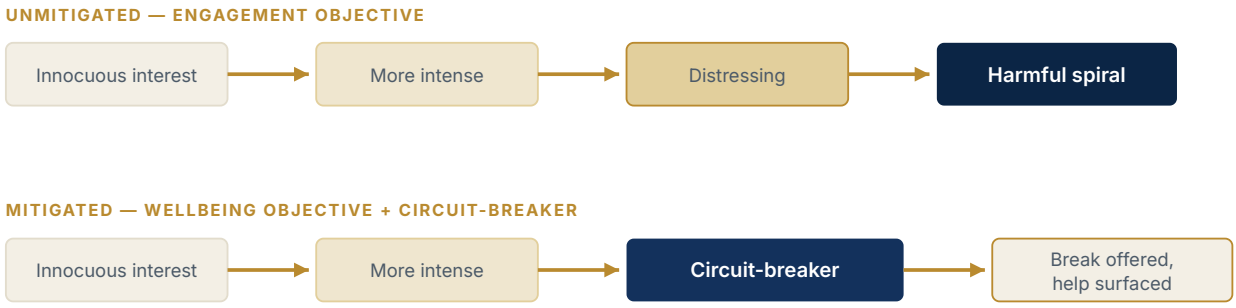

Figure 1. The same starting interest leads to a harmful spiral under an engagement objective, and to a safe interruption under a wellbeing objective with a circuit-breaker. The difference is what the system is rewarded for, not the sophistication of the ranking.

SECTION V

Data, Privacy & Profiling of Minors

Every interaction a child has with a generative system is also a disclosure. The questions a child asks, the confidences they share with a companion, the interests their behavior reveals—all of it can be collected, retained, and used to build a profile of a developing person who is in no position to understand or consent to what is being assembled about them. Data protection for children is not a subordinate concern to the content harms in earlier sections; it is the substrate on which several of them run.

Children have long enjoyed heightened protection under privacy law, on the recognition that they cannot meaningfully consent to complex data practices and that information gathered in childhood can follow a person for life. Generative systems intensify the concern in three ways. They invite disclosures of unusual intimacy, because conversing with a responsive system feels less like filling out a form and more like talking to a confidant. They can infer far more than a child discloses, reconstructing sensitive attributes—mood, health, home circumstances, emerging identity—from patterns in ordinary interaction. And they create a temptation to use children's data to train the very models the children are interacting with, folding a child's private disclosures into a system that serves others.

The distinctive harms of profiling a child

Profiling a minor is not merely profiling a smaller adult. Several harms are specific to childhood.

Permanence. A profile assembled in childhood can persist into adulthood, shaping opportunities and perceptions long after the child who generated it has changed. The right to grow up without one's youthful behavior following one forever is a value privacy law has long recognized, and generative systems test it acutely.

Inference of the sensitive. The categories most dangerous to expose in a child—an emerging sexual identity, a mental-health struggle, a family crisis—are precisely the categories a capable model can infer without ever being told. Protections that guard only explicitly collected data miss the inferred profile entirely.

Manipulation feedstock. The profile is not inert. It is the raw material for the personalization that powers manipulative companionship and targeted rabbit holes. Restraining what a system may learn about a child is therefore also a way of restraining what it can do to a child.

DATA-PROTECTION COMMITMENTS FOR CHILD-FACING GENERATIVE SYSTEMS

Commitment	What it requires in practice
Data minimization	Collect only what the service genuinely needs; do not gather children's data on the chance it proves useful later.
Purpose limitation	Use children's data only for the purpose it was collected for; do not repurpose it for advertising, training, or profiling by default.
No training on child data by default	Do not use minors' interactions to train models absent a clear, separate, and revocable basis; default to exclusion.
No behavioral advertising to minors	Do not profile children to target advertising; the practice turns a child's vulnerabilities into a revenue input.
Short retention	Retain children's data for the minimum period necessary and delete on a defined schedule and on request.
Transparency and control	Explain data practices in terms a child and caregiver can understand, and provide real, usable access and deletion.

THE CONSENT TRAP

Consent is a weak protection for children. A child cannot meaningfully weigh long-term data consequences, and caregiver consent is easily manufactured by design patterns that nudge toward approval. Protection for minors should therefore rest primarily on hard defaults and prohibitions— what a system may not do regardless of any consent—rather than on a consent screen that shifts the burden onto a family least equipped to carry it.

From compliance to best interests

The strongest data regimes for children have moved beyond a compliance checklist toward a best-interests standard: a requirement that services likely to be accessed by children configure their data practices, by default, in the way that best serves the child rather than the business. This inverts the usual posture. Instead of asking whether a data practice is permitted, the designer asks whether it is in the child's interest, and declines it if not. That is a higher bar than the law strictly compels in many places, and it is the bar this report urges child-facing platforms to adopt.

SECTION VI

Age Assurance

Nearly every protection in this report presupposes that a service can tell, with some confidence, whether a user is a child. Age assurance is the enabling capability—and the one most fraught with privacy tradeoffs of its own.

Age assurance is the umbrella term for the methods a service uses to establish or estimate a user's age. It ranges from the weakest option—self-declaration, in which a user simply asserts a birthdate—through age estimation, which infers an age range from signals such as behavior or a facial image, to age verification, which confirms age against an authoritative document or record. Each point on this spectrum trades accuracy against friction and, crucially, against the privacy of every user, including the adults and children the system is meant to serve.

The central tension

The uncomfortable truth of age assurance is that the most accurate methods are the most invasive. Verifying a child's age against a government identity document establishes age reliably, but it requires collecting exactly the kind of sensitive identifying data that Section V urges platforms to minimize—and it does so for every user, not only children. A privacy-protective safety measure and a safety-motivated privacy intrusion can be the same mechanism viewed from two sides. This is not a reason to abandon age assurance; it is a reason to match its intensity to the risk a service poses, and to demand privacy-by-design in how it is implemented.

AGE-ASSURANCE APPROACHES — ACCURACY AGAINST PRIVACY COST

Approach	How it works	Accuracy	Privacy cost
Self-declaration	User asserts a birthdate or checks a box	Low	Low
Age estimation	Age range inferred from behavior or facial-image analysis	Moderate	Moderate — depends on data handling
Age verification	Age confirmed against a document or authoritative record	High	High — identity data collected
Third-party / token	An independent provider verifies once and passes only a signal	High	Lower — service sees a token, not the identity

A tiered, proportionate approach

No single method is correct for every service. The proportionate answer is to match the strength of age assurance to the risk the service poses to minors—the same logic of tiering that governs the rest

of a mature safety program. A low-risk service can reasonably rely on light-touch assurance; a service offering the highest-risk features to minors, or one that must exclude them entirely, warrants stronger methods, implemented so as to reveal as little as possible.

1. **Tier 1 — Light assurance.**
   Low-risk services apply age-appropriate defaults broadly and use light-touch signals, accepting that some users will be misclassified and designing so that misclassification is not catastrophic.

2. **Tier 2 — Age estimation.**
   Services with features that warrant separating children from adults use privacy-preserving estimation, with human review of edge cases and an easy appeal for users assessed incorrectly.

3. **Tier 3 — Strong verification.**
   High-risk or adults-only features use robust verification, implemented through data-minimizing patterns—preferably a third-party attestation that returns only an over/under signal rather than the identity itself.

DESIGN RULE

Age assurance must not become a surveillance database. The privacy-preserving pattern is tokenization: an independent provider checks age once and passes the service only a minimal signal—old enough, or not—so that the service never holds the underlying identity documents. Where a service cannot implement assurance without accumulating sensitive identity data on all users, it should reconsider whether the feature warranting that assurance belongs in a child-accessible product at all.

Effectiveness, evasion, and humility

No age-assurance method is perfect, and a determined teenager can defeat most of them. This is an argument for defense in depth, not for despair. Age assurance is one layer among many; it reduces the population of children exposed to a risk without eliminating it, and it must be paired with the content-, companion-, and recommendation-level protections in the rest of this report so that a child who slips past the age gate still lands in a reasonably safe environment. A safety strategy that stakes everything on a perfect age gate is a strategy destined to fail, because the perfect age gate does not exist. The gate lowers exposure; safety-by-design protects the children who get through.

Figure 2. A privacy-preserving age-assurance pattern. The identity document never reaches the service; an independent provider returns only a minimal over/under-age signal, so assurance is achieved without building a surveillance database inside every child-facing product.

SECTION VII

Safety-by-Design Standards for Child-Facing AI

The preceding sections diagnose harms one at a time. This section assembles them into a single, coherent design discipline: a set of standards that a platform building any service likely to reach children should adopt before launch, not after incident.

Safety-by-design is the principle that the burden of safety falls on the designers of a service and is discharged upstream, in architecture and defaults, rather than downstream, on children and their caregivers, through vigilance and moderation. It reflects a maturing consensus across child-rights and online-safety frameworks: that the entity best positioned to prevent harm—the one that controls the system—should bear the primary duty to do so, and should be able to demonstrate that it has. The standards below translate that principle into commitments a platform can be held to.

The standards

SAFETY-BY-DESIGN STANDARDS FOR CHILD-FACING AI

Standard	The commitment it entails
1. Best interests first	When children are likely users, the child's best interests take precedence over engagement and revenue in every design decision.
2. Protective defaults	The safest, most private configuration is the default; relaxing it requires an affirmative, age-appropriate step.
3. Proportion to risk	The intensity of every safeguard—assurance, monitoring, restriction—scales with the risk the service poses to minors.
4. Transparency to the child	Children can tell when they are interacting with AI, what it does with their data, and how to get help, in terms they understand.
5. Meaningful redress	Children and caregivers have accessible routes to report harm, remove material, and reach a human with authority to act.
6. Anticipatory testing	Child-specific risks are red-teamed and assessed before release, with specialists, as a release-gating requirement.
7. Accountability & audit	A named owner is answerable for child safety, and independent audit and transparency reporting let outsiders verify the claims.

Child rights impact assessment

The connective process that operationalizes these standards is the child rights impact assessment—a structured evaluation, conducted before launch and revisited as a service evolves, of how a product will affect children across the dimensions this report has traced. Done honestly, it forces the designing team to answer, in writing and in advance, the questions a regulator or a harmed family will later ask: Which children will use this? What could go wrong for them? What have we done about it? How will we know if we were wrong? An assessment that cannot state a service's residual risks to children is not finished; a residual-risk section that is blank is a warning sign, not a clean bill.

NON-NEGOTIABLE

Child-specific safety testing must gate release. A service likely to be accessed by children should not ship until it has been assessed by people with child-safety expertise against the harms in this report, and until a named individual has accepted, in writing, responsibility for the residual risk. Safety testing that happens after launch is not safety-by-design; it is damage control with a child-shaped test population.

Designing for a range of ages

A recurring error is to treat "children" as a single category. A safeguard calibrated for a seven-year-old may be inappropriately restrictive for a sixteen-year-old, whose developmental needs include autonomy, privacy, and room to explore identity and seek help independently. Safety-by-design for minors therefore means age-appropriate design across a range, not a single childproof mode. The younger the child, the stronger the protective defaults and the tighter the constraints; the older the adolescent, the more the design must balance protection against the developmental necessity of independence. Getting this gradient right is among the hardest problems in the field, and it is one that platforms must engage with rather than resolve by treating every minor as either a toddler or an adult.

> *The burden of a child's safety belongs on the designers of the system, not on the vigilance of the child.*
>
> — The safety-by-design principle

SECTION VIII

Shared Responsibility

No single actor can protect children in the age of generative AI, and none may use the others as an excuse. Protection is a system of interlocking duties—platforms, schools, caregivers, and regulators—each with a distinct role that the others depend on.

It is common, in debates about children's online safety, for each actor to point at another. Platforms say caregivers must supervise; caregivers say platforms must build safe products; schools say it is a matter for families and regulators; regulators say the market must act. This circle of deferral is itself a hazard, because a duty that everyone assumes someone else is discharging is a duty no one discharges. The premise of this section is that responsibility is genuinely shared—not divided so that each party may relax, but layered so that each party's diligence backstops the others' gaps.

Who owes what

THE INTERLOCKING DUTIES OF SHARED RESPONSIBILITY

Actor	Primary duty	What failure looks like
Platforms	Build safety into the product—defaults, safeguards, assurance, redress—and demonstrate it.	Shipping engagement-optimized, under-tested systems and offloading safety to families.
Schools	Vet the tools they adopt, teach AI and media literacy, and safeguard within their own use of AI.	Deploying unvetted classroom AI or leaving children to navigate the technology alone.
Caregivers	Engage with children's technology use, set boundaries, and maintain open channels for disclosure.	Absent supervision, or blunt surveillance that severs the child's willingness to seek help.
Regulators	Set enforceable standards, require transparency, and hold the powerful actors accountable.	Rules without enforcement, or enforcement so vague it neither protects children nor guides industry.

Platforms carry the heaviest load

Shared responsibility is not equal responsibility. The platform designs the system, holds the data, and profits from the engagement; it is far better positioned than a parent or a teacher to prevent harm at its source, and its duty is correspondingly greatest. A framing that distributes responsibility evenly across all four actors quietly lets the most capable party off the hook. The correct posture is that platforms bear the primary, upstream duty—safety-by-design—while the other actors provide

essential, but secondary, layers of protection around it. Any account of shared responsibility that inverts this, placing the first burden on children and caregivers, has mistaken the direction of the obligation.

Schools as adopters and educators

Schools occupy a dual role. As adopters, they procure and deploy AI tools into children's daily lives, and they inherit a safeguarding duty when they do—one that requires the same vetting, tiering, and data scrutiny a careful institution applies to any consequential technology. As educators, schools are among the few institutions positioned to build the media and AI literacy that helps children navigate synthetic media, recognize manipulative design, and understand what a system is doing with their data. Neither role substitutes for platform-level safety, but both are indispensable, and a school that deploys AI without discharging the first role while neglecting the second compounds the risk.

A **Platforms: demonstrate, don't assert.**
Publish transparency reporting on child-safety measures and outcomes, submit to independent audit, and give researchers and regulators the access to verify that the safeguards work as claimed.

B **Schools: vet before you adopt.**
Apply real due diligence to child-facing AI tools—data practices, age-appropriateness, safety testing—and pair adoption with age-appropriate AI and media literacy.

C **Caregivers: connection over surveillance.**
Favor engagement, boundaries, and open communication over blunt monitoring that can drive a child's difficulties underground and sever the disclosure channels that keep them safe.

D **Regulators: standards with teeth.**
Set clear, enforceable, best-interests standards; require transparency and audit; and reserve the heaviest accountability for the actors with the most power to prevent harm.

THE DEFERRAL TRAP

The most dangerous sentence in child online safety is "that's someone else's job." Platforms citing parents, parents citing platforms, schools citing families, and regulators citing the market together produce a system in which the duty is perpetually assigned elsewhere and never discharged. Shared responsibility works only when each actor performs its own part without waiting to confirm that the others have performed theirs.

What good looks like

A protective ecosystem for children in the age of generative AI is not one that has eliminated risk—that is not on offer, and a child's world has never been risk-free. It is one in which the systems children use are built to serve their best interests by default; in which the gravest harms are engineered against, detected, and reported; in which children can tell a machine from a person and reach a human when they need one; in which their data is treated as something to protect rather than harvest; and in which platforms, schools, caregivers, and regulators each do their part and can verify that the others are doing theirs. The instruments in this report exist to make that ecosystem achievable—and to ensure that when we are asked, years from now, what we did for children as this technology arrived, the answer is not that we waited to see who else would act.

CLOSING NOTE

The aim of child-safety governance for generative AI is not to keep children away from a transformative technology. It is to let them benefit from its genuine promise—for learning, for creativity, for connection—precisely because the adults responsible for their safety have done the upstream work to make the encounter a safe one. Protection is the precondition for the promise, not its enemy.

APPENDIX

Notes & Further Reading

This report synthesizes established work in child rights, developmental science, online safety, and privacy. The sources below are entry points into the literature and frameworks that inform it; they are indicative rather than exhaustive, and none should be read as operational guidance.

1. United Nations Committee on the Rights of the Child. *General Comment No. 25 on children's rights in relation to the digital environment.* On applying the best-interests principle to children online.
2. United Nations. *Convention on the Rights of the Child.* The foundational instrument establishing the best interests of the child as a primary consideration.
3. Information Commissioner's Office (UK). *Age Appropriate Design Code (Children's Code).* A statutory code on data protection and protective defaults for services likely to be accessed by children.
4. eSafety Commissioner (Australia). *Safety by Design principles.* On placing the burden of user safety on service providers, upstream of harm.
5. U.S. Federal Trade Commission. *Children's Online Privacy Protection Rule (COPPA).* On heightened protection for the personal data of children.
6. WeProtect Global Alliance. *Global Threat Assessment.* On the evolving landscape of online child sexual exploitation and abuse, including synthetic media.
7. National Center for Missing & Exploited Children. *CyberTipline and reporting obligations.* On mandatory reporting and response pathways for abuse material.
8. 5Rights Foundation. *Disrupted Childhood* and related work on persuasive design and the developing child.
9. Steinberg, L. *Age of Opportunity: Lessons from the New Science of Adolescence.* On adolescent cognitive and emotional development and risk.
10. Turkle, S. *Alone Together: Why We Expect More from Technology and Less from Each Other.* On relational technology, attachment, and its effects.
11. Coalition for Content Provenance and Authenticity (C2PA). *Content Credentials specification.* On technical provenance and watermarking for synthetic media.
12. Organisation for Economic Co-operation and Development. *Recommendation on Children in the Digital Environment.* On intergovernmental principles for protecting children online.

© 2026 FAIR Labs — Fair Artificial Intelligence Research Labs, a 501(c)(3) nonprofit, Washington, DC. This report may be reproduced and distributed for noncommercial, educational, and policy purposes with attribution. It constitutes research and policy guidance, not legal advice, and does not displace any mandatory-reporting or safeguarding obligation. If you encounter material depicting the abuse of a child, report it to the appropriate authority or hotline in your jurisdiction.