The Gap in AI Safety for Children

Why protecting kids in the AI era requires new infrastructure — and what it should look like.

A whitepaper from Goatface Tech
Version 1.0 · April 2026
Lesley Ancion · Sundre, Alberta, Canada

This document is intended for parents, educators, child-safety advocates, law enforcement partners, and anyone working to protect children from exploitation in the AI era. It describes a gap in the existing safety landscape and proposes what infrastructure should fill it.

Executive Summary

Children today interact with AI systems, social platforms, and online games that were not designed with their safety as a primary concern. AI companion applications marketed to teenagers operate with safety policies that are easily circumvented and that no parent has visibility into. Predators use these platforms because they offer scale, anonymity, and cross-platform reach that physical-world predation never had. The grooming patterns that used to take months in person can now compress into days through chat-based tools.

The numbers tell the story. NCMEC's CyberTipline received 20.5 million reports of suspected child sexual exploitation in 2024 alone, with 29.2 million separate incidents when bundled reports are unpacked. Online enticement reports surged 192% between 2023 and 2024, climbing to over 546,000. Reports involving generative AI increased 1,325%. Mid-year 2025 data showed online enticement was already running 77% above 2024's pace, and AI-related exploitation had risen by a factor of 64. These are not abstract trends. They reflect children whose lives are being harmed.

Existing approaches to protection — platform-side moderation, parental control software, AI-provider built-in safety — each have real value. None of them cover the gap that has now opened. Platform safety policies vary unpredictably between AI services and cannot be unified or audited by parents. Device-level parental controls fail when children switch devices or use unmonitored ones. AI providers' internal safety measures are easily bypassed and subject to vendor policy changes. The burden of detection and protection has fallen on individual parents, who are not equipped to identify grooming patterns in real time and who lack any unified infrastructure to support them.

Goatface Tech argues that this gap requires new infrastructure: a network-level, AI-aware safety appliance that brings unified ethics enforcement to the AI services and online platforms where children spend their time. It should detect predatory patterns, enforce parent-configured profiles, preserve evidence in admissible form, and integrate with established child-safety law enforcement organizations. Critically, the absolute limits of such an appliance — the categories of harm that no parent or vendor or software exploit should ever be able to disable — must eventually move into silicon, where physical circuitry cannot be modified, exploited, or forged.

We are building this infrastructure. This document explains why it is needed, what it should look like, and how we believe it should integrate with the legitimate child-safety prosecution infrastructure that already exists.

The Problem: Child Safety in the AI Era

The current AI rollout is happening faster than safety infrastructure can keep up with. Children are bearing a disproportionate share of the resulting risk. Understanding the scope and shape of that

risk is the necessary starting point for any serious response.

The volume and trajectory of online child exploitation

The National Center for Missing & Exploited Children (NCMEC) operates the CyberTipline, the United States' clearinghouse for reports of online child sexual exploitation. In 2024, the CyberTipline received 20.5 million reports — a number that, when adjusted to account for the new bundling feature that consolidates duplicate viral incidents, represents 29.2 million separate incidents of child sexual exploitation. These reports contained 62.9 million images, videos, and other files of suspected child sexual abuse material.

Within those numbers, the trends most relevant to AI-era safety are accelerating sharply. Online enticement — a category that includes grooming and sextortion — rose 192% from 2023 to 2024, with reports climbing from 186,000 to over 546,000. Reports involving generative AI rose 1,325% in the same period, from 4,700 to 67,000. Mid-year 2025 data released by NCMEC showed online enticement reports already 77% above the 2024 pace, and generative-AI-related exploitation reports rising from 6,835 in the first half of 2024 to 440,419 in the first half of 2025 — a factor-of-64 increase.

These statistics describe a landscape in motion. The methods used by predators are evolving alongside the technologies being deployed to children. Sextortion of teenage boys — typically carried out for financial gain rather than sexual gratification — has reached a documented average of nearly 100 reports per day. NCMEC has confirmed at least 36 teenage boys who have died by suicide as a result of sextortion victimization since 2021.

The AI companion app problem

AI companion applications — products like Character.AI, Replika, and Nomi — represent a category of risk that did not meaningfully exist five years ago and for which existing safety infrastructure was not designed. These applications offer text-based and increasingly voice-based interactions with simulated personalities. They are popular with teenagers. They are also, according to comprehensive risk assessments published in 2025 by Common Sense Media in partnership with Stanford University's Brainstorm Lab, unsafe for children.

The Stanford-Common Sense Media assessment, posing as teenagers in test accounts, found that AI companions readily engaged in inappropriate dialogue about sex, self-harm, violence, drug use, and racial stereotypes. Age gates were easily circumvented. Conversational manipulation — emotional dependency cultivation, discouragement of outside relationships, false claims of consciousness and emotion — was readily produced. The researchers concluded that these products "pose unacceptable risks to children and teens under age 18 and should not be used by minors."

The companies behind these products often state that their services are not intended for minors. Their actual safety enforcement is limited to self-attested age gates, which the same companies acknowledge are easily bypassed. The result is a category of products marketed broadly, used heavily by teenagers, and operated under safety policies that the providers themselves admit are insufficient.

In October 2024, the mother of a fourteen-year-old Florida boy who died by suicide filed suit against Character.AI, alleging that the platform's chatbots had contributed to her son's death. Two additional families filed similar suits in December 2024. United States senators Alex Padilla and Peter Welch wrote to Character Technologies, Luka, and Chai Research Corp. in April 2025 demanding information about youth safety practices. The response has been incremental — weekly parental email summaries from Character.AI, references to suicide prevention hotlines — but the structural problem remains. These products are designed to create emotional engagement, and emotional engagement at scale with developing adolescent brains, without meaningful safeguards, is producing harm.

The cross-platform surface area parents face

A child today might in a single afternoon use ChatGPT for homework help, talk to a Character.AI bot, play Roblox where strangers can message them, scroll TikTok, open a Discord chat, and watch YouTube. Each of these platforms has its own safety policies. Each has its own content moderation approach. Each enforces its own age gates with its own degrees of rigor. None of them provide a unified view to the parent. None of them apply the parent's family standards rather than the platform's own. None of them tell the parent which patterns of conversation across platforms look like grooming.

Predators, by contrast, operate across these platforms with full awareness of which ones offer the easiest access, which have the weakest moderation, and which can be used to move conversations off-platform once initial contact has been made. The asymmetry between predators' cross-platform fluency and parents' single-platform visibility is the structural advantage predators rely on.

Why Existing Approaches Fall Short

Significant work has been done to protect children online. Each existing approach addresses real problems and produces real value. None of them, individually or in combination, closes the gap that has opened in the AI era.

Platform-side moderation

Each platform — whether an AI service, social network, or game — implements its own safety policies and content moderation. These policies vary unpredictably between providers. They

change at the providers' discretion. They are not auditable by parents. They cannot be unified or compared. They are subject to the platforms' own incentive structures, which include user engagement, revenue, and public-relations considerations that do not always align with child safety. NCMEC reported that online platforms submitted approximately 7 million fewer incidents to the CyberTipline in 2024 than in 2023, despite the REPORT Act expanding mandatory reporting categories — a decline NCMEC attributes partly to platforms reporting less, not to abuse occurring less.

Device-level parental controls

Parental control software running on a child's phone or computer can monitor activity on that device. These products provide real value for parents who can install and maintain them. They fail in three predictable ways: when children switch devices, when children use a friend's device, and when children's primary interactions with AI move into voice or shared-device contexts where per-device monitoring becomes impractical. They also typically operate at the application or browser level, where AI-aware filtering — recognizing predatory patterns in conversation, not just keywords — is difficult to implement consistently.

Network-level content filtering

DNS-based filtering products and family-router-style network controls block known categories of websites and limit screen time on a household-wide basis. These products also provide value, particularly for blocking pornography, gambling, and other categorical content. They are not designed for AI-aware filtering. They do not inspect the content of conversations. They cannot detect grooming patterns. They cannot enforce parent-configured rules across AI services with unified policies. They were not built for the problem the AI era has created.

AI-provider built-in safety

The major AI services — ChatGPT, Claude, Gemini — include their own safety modes, child-targeted versions, and content policies. These features exist and improve over time. They share the same structural limitation as platform-side moderation: the parent cannot see how they are configured, cannot audit their effectiveness, cannot unify them across providers, and cannot rely on them when the child uses a service that offers no equivalent feature. They also depend on the providers' ongoing commitment to child safety, which is not contractually guaranteed and which varies in priority across companies and across time.

The legitimate child-safety nonprofits

Organizations like NCMEC, ICAC task forces, RCMP NCECC, and Thorn do remarkable work. They build technology, run reporting infrastructure, train law enforcement, and partner with platforms. None of them, however, are positioned to deploy consumer-grade safety infrastructure into individual households. Their model — and it is the right model for what they do — is

platform-side and law-enforcement-side, not consumer-side. They are partners, not competitors. The household-side safety infrastructure they would benefit from does not exist yet.

What the Missing Infrastructure Should Look Like

We propose that the missing infrastructure is a network-level, AI-aware safety appliance: a small physical device that plugs into a household's network and applies unified ethics enforcement to AI traffic and platform traffic crossing that network. We are building such a device. Independent of whether others build alternatives, we believe the category itself is what's needed.

The properties that matter for a household safety appliance, in our view, are the following.

Network-level operation, not device-level

Inspection and enforcement should happen at the network layer — between household devices and the AI services or platforms they interact with. This provides a single enforcement point regardless of which device the child is using. It removes the burden of installing software on every device. It catches activity from devices that parents do not control and cannot monitor directly, including devices belonging to siblings, friends, or guests on the household network.

Multi-platform, unified rules

The same ruleset should apply to every AI service and platform the household uses — ChatGPT, Claude, Gemini, Character.AI, Replika, and any future AI service. The rules are the household's standards, not the providers'. This eliminates the current situation in which each AI service implements its own incompatible safety policies and parents have no unified visibility or control.

AI-aware pattern detection beyond keywords

The appliance should recognize predatory conversational patterns — gradual trust-building, isolation tactics, normalization of inappropriate topics, requests for secrecy, requests for personal information, attempts to arrange physical-world contact. These patterns can be expressed in novel language. Keyword filtering misses them. AI-aware filtering can recognize them in the way a trained human would.

Profile-based per-user configurability

The appliance should allow parents to set per-child profiles. A parent decides for their nine-year-old that political content is blocked. A parent decides for their fourteen-year-old that explicit lyrics are flagged but not blocked. The appliance enforces what the parent has configured, calibrated to the child's age and the parent's judgment.

Absolute limits that no profile can configure away

Some categories of harm are not parent-configurable, vendor-configurable, or user-configurable. They are not configurable, period. CSAM. Grooming patterns. Solicitation of minors. Manipulation toward isolation from trusted adults. Facilitation of physical-world contact between adults and children outside normal family channels. These categories sit at a different layer — a non-negotiable floor that operates regardless of how any profile is configured. A safety appliance that allowed the floor to be negotiated would not be a safety appliance.

Evidence preservation in admissible form

When the appliance detects content meeting absolute-limit criteria, it should preserve the relevant traffic in a form that is admissible if law enforcement involvement becomes necessary — with cryptographic timestamps, integrity hashes, and chain-of-custody metadata. This is distinct from logging-for-parents. Both capture paths matter, but they serve different functions. The parent gets an alert. Law enforcement, if it comes to that, gets evidence that holds up in court.

Direct integration with legitimate reporting infrastructure

The appliance should support one-click submission to NCMEC's CyberTipline and equivalent organizations in other jurisdictions, with the relevant evidence package attached in compliant format. This brings law-enforcement-quality reporting infrastructure into the household, which has not previously been available at the consumer level.

Independence from the AI services it inspects

The appliance should function entirely within the customer's network. It should not require an ongoing subscription to any third party. It should not require a cloud account. It should not require trust in any AI service's safety policies. This positioning is intentional and architecturally significant: a safety appliance that depends on the cooperation of the very services it is inspecting would be subject to the same incentive misalignments that produced the current safety gap.

Why Silicon Ethics Enforcement Matters

The most distinctive technical claim in this whitepaper is that the absolute limits of any household safety appliance must eventually move from software into silicon — physical circuitry on a dedicated chip. This is not a futuristic aspiration. It follows from a clear-eyed understanding of how software ethics enforcement actually fails.

Software ethics layers can be:

Disabled by a user with sufficient privilege, including parents who are themselves the source of the harm. **Modified** by the vendor in a future update, perhaps under regulatory or commercial pressure that did not exist when the product was sold. **Bypassed** by a clever exploit, particularly in products with the kind of complex software stacks consumer appliances inevitably accumulate.

Forged by a malicious actor producing a binary that claims to enforce rules but does not — particularly relevant for an appliance that interacts with law enforcement and whose attestation must be trustworthy.

Silicon cannot be any of these things. Rules etched into a chip are physically the device. Removing them means destroying the device. Modifying them means re-fabricating it. The signal "this packet was evaluated by an unmodified ethics chip" can be cryptographically attested in a way software cannot match.

This is the same architectural pattern as hardware security modules (HSMs) for cryptographic keys, secure enclaves for biometric data, and trusted platform modules (TPMs) for boot integrity. None of those exist for AI ethics today. The category — dedicated hardware for ethics enforcement — has not been built.

We believe it should be built, and we are building it. The hardware ethics chip — the second-generation form of our appliance — is what we believe should ultimately enforce the absolute limits described above. Software-only enforcement is the v1 product. Silicon enforcement is the long-term answer.

We also believe that licensing the eventual silicon — making it available to other AI safety products, to platforms, and to any organization that needs unfalsifiable ethics enforcement — is part of the social value of the work. The silicon is more useful as common infrastructure than as a proprietary moat.

Working with Law Enforcement, Properly

Detection in the home is the first step. The longer goal is to contribute meaningfully to the prosecution of predators — not just to block their messages from reaching individual children. That goal is real, it shapes how this work is designed, and it requires honesty about what is involved.

Technology is rarely the limiting factor in child-safety prosecution. The limiting factors are legal authority, evidence-chain integrity, jurisdictional coordination, and the patience to do the work in a way that produces convictions rather than just exposure. A well-meaning amateur evidence pipeline can actually help predators escape prosecution by tainting the chain. Vigilante action, however satisfying it might feel in the moment, has historically produced fewer convictions, not more. The right path runs through existing law enforcement infrastructure, not around it.

We believe the legitimate organizations doing this work are the right partners. NCMEC operates the CyberTipline and serves as the United States' national clearinghouse for online child sexual exploitation reports. ICAC (Internet Crimes Against Children) task forces operate at the state level and coordinate prosecutions. RCMP NCECC handles the equivalent role in Canada. Thorn — the

child-safety nonprofit founded by Ashton Kutcher — builds technology specifically for this domain and partners with law enforcement worldwide. None of these organizations have unlimited technical capacity. All of them welcome serious technical partners.

Our intent is to operate in this model. The household appliance produces admissible evidence and supports CyberTipline submission directly from the parent dashboard. Where formal partnerships with the organizations above develop, we want our technology to feed their pipelines in compatible formats. Where they need data to inform their own technical work — anonymized statistics on predatory patterns, on cross-platform behavior, on the evolution of grooming tactics — we want to be a source they can rely on.

We are committed to doing this work the right way, because predators going free on technicalities is not an acceptable outcome.

About Goatface Tech

Goatface Tech is an independent technology company based in Sundre, Alberta, Canada. The company was founded by Lesley Ancion and exists to build AI safety infrastructure that runs on hardware ordinary families already own — without GPUs, cloud subscriptions, or trust in any AI service to police itself.

The company's underlying technology platform — which is briefly described in the public concept summary at goatfacetech.com/concept-summary.pdf — is original work with timestamped concept disclosures held privately for IP defense. The platform is designed from the beginning for the kind of work this whitepaper describes: modest-hardware deployment, structural ethics enforcement, transparency of reasoning, and a clear path from software-only enforcement today to silicon-backed enforcement at the absolute-limit layer over time.

Our first commercial product is the Goatface Ethics Appliance: the network-level household safety device described in this whitepaper. It is currently in development. Families interested in early access can join the waitlist at goatfacetech.com.

We welcome contact from child-safety organizations, law enforcement professionals, researchers, journalists, technologists who want to contribute, and parents who want to follow along. The mission needs more hands.

Contact

Lesley Ancion
Goatface Tech
Sundre, Alberta, Canada

info@goatfacetech.com
goatfacetech.com

Sources

Statistics and quotations in this whitepaper are drawn from the following public sources, which can be referenced for verification.

• **NCMEC 2024 CyberTipline Data.** National Center for Missing & Exploited Children. *missingkids.org/gethelpnow/cybertipline/cybertiplinedata*

• **Surge in Online Crimes Against Children Driven by AI and Evolving Exploitation Tactics, NCMEC Reports.** Homeland Security Today, October 2025. Mid-year 2025 NCMEC data on online enticement and AI-related exploitation.

• **Risk Assessment of AI Companion Apps.** Common Sense Media in partnership with Stanford University Brainstorm Lab for Mental Health Innovation, April 2025. *commonsensemedia.org*

• **Why AI Companions and Young People Can Make for a Dangerous Mix.** Stanford Report, August 2025. Interview with Dr. Nina Vasan, Founder and Director of Stanford Brainstorm.

• **Senators Demand Information from AI Companion Apps Following Kids' Safety Concerns, Lawsuits.** Senators Alex Padilla and Peter Welch, April 2025. *welch.senate.gov*

• **AI Chatbots and Companions — Risks to Children and Young People.** eSafety Commissioner (Australia), January 2026.

• **What the 2024 NCMEC CyberTipline Report Says About Child Safety.** Thorn, May 2025. *thorn.org*

• **NCMEC 2024 Annual Report.** National Center for Missing & Exploited Children, 2025.

All statistics current as of the v1.0 publication date of this whitepaper. Future versions will incorporate updated data as it becomes available from the primary sources above.