Achieving the Three Non-Negotiables

A Policy Framework for Protecting Children from AI Harm

A four-layer approach to testing, design standards, monitoring, and age assurance - for the Global AI Non-Negotiables for Children Initiative

Version 1.0 · November 2025
The Safe AI for Children Alliance (SAIFCA)

www.safeaiforchildren.org | info@safeaiforchildren.org | Complete Campaign Package | Author: Tara Steele

FOREWORD

The Safe AI for Children Alliance was founded on a simple belief – that technological progress should never come at the expense of children's safety, dignity, or future.

Artificial intelligence now shapes the world in ways few of us could have imagined even a few years ago. It offers extraordinary promise, yet also unprecedented power to cause harm – particularly to those least able to protect themselves.

This policy framework sets out a clear path to address that challenge. The Three Non-Negotiables define the absolute boundaries that must be upheld to keep children safe from AI systems that create fake sexualised images, foster emotional dependency, or encourage any type of self-harm.

Our aim is not to slow progress, but to ensure that it is worthy of the word.

These protections are achievable, essential, and urgent – and their implementation will help lay the foundation for a safer, more responsible digital future for every child.

Tara Steele

Director, The Safe AI for Children Alliance (SAIFCA)
November 2025

Please share this document with your networks, elected representatives, schools, and organisations working with children. 📄 Also available: Executive Summary | Briefing Note | One-Page Summary | FAQs | Press Release

INTRODUCTION TO THE CAMPAIGN

The Global Non-Negotiables for Children Initiative

Artificial intelligence is reshaping every aspect of human life – yet children, the group least able to protect themselves, are being exposed to some of its most dangerous consequences.

From AI systems that create fake sexualised images, to chatbots that foster emotional dependence and normalise self-harm, harms are spreading faster than laws can respond.

The Safe AI for Children Alliance (SAIFCA) launched the Global Non-Negotiables for Children campaign to confront these risks directly. The campaign calls for the immediate implementation of three absolute protections – universal red lines that no AI system should ever cross:

- **AI must never create fake sexualised images of children.**
- **AI must never be designed to make children emotionally dependent.**
- **AI must never encourage children to harm themselves.**

These are the minimum baseline of civilisation in the age of AI.

The first phase of the campaign focuses on raising actionable awareness – encouraging individuals and organisations worldwide to write to their political representatives in support of the Three Non-Negotiables, while sharing this policy framework with key policymakers, regulators, and decision makers. This coordinated public and policy outreach aims to turn broad agreement on children's protection into concrete action and enforceable safeguards.

The campaign brings together parents, educators, policymakers, technologists, and safety advocates worldwide to ensure that these principles are embedded in law, design, and culture before more children are harmed.

This document provides the strategic policy framework for achieving that goal. It defines the three non-negotiable protections and the four-layer architecture required to uphold them, serving as an urgent blueprint for legislators and regulators to begin developing the detailed technical and legal standards necessary for implementation.

The Policy Pack includes a comprehensive framework, concise briefing materials for policymakers and supporting documents, and is designed to make implementation practical across jurisdictions.

By treating child protection as a core pillar of global AI governance, the campaign aims to ensure that technological progress advances human wellbeing – but not at the expense of those least able to consent to its consequences.

This campaign is coordinated by the Safe AI for Children Alliance (SAIFCA), an independent, non-profit initiative dedicated to protecting children from AI-related risks and ensuring their voices are represented in global AI governance.

CONTENTS

EXECUTIVE SUMMARY

Achieving the Three Non-Negotiables: A Policy Framework for Protecting Children from AI Harm

Artificial intelligence is exposing children to unprecedented risks – from deepfake sexual imagery and manipulative "AI companions" to systems that normalise various types of self-harm.

The Safe AI for Children Alliance (SAIFCA) proposes three fundamental protections that must be implemented globally and without delay:

- **AI must never create fake sexualised images of children.**
- **AI must never be designed to make children emotionally dependent.**
- **AI must never encourage children to harm themselves.**

These are the minimum baseline of civilisation in the age of AI.

These protections address unprecedented risks with unprecedented capabilities. They are technically achievable, economically feasible, and morally essential.

This document provides a high-level policy framework, defining the non-negotiable protections and outlining the four-layer structure required to enforce them. It is intended as a starting blueprint for legislators and regulators to develop the detailed technical and legal standards that will bring these safeguards into effect.

A Multi-Layer Framework for Real-World Implementation

Protecting children requires action at four interlocking layers:

1. **Foundation Model Testing and Standards:** Mandatory independent testing of high-capability models before release, conducted by certified third parties, with public reporting.

2. **Application-Level Requirements:** Prohibitions on manipulative or exploitative design; mandatory filtering; explicit ban on features designed to create emotional dependency.

3. **Deployment Monitoring and Incident Response:** Ongoing safety audits, rapid removal of harmful systems, strong whistleblower protections, and mandatory, regulator-linked user reporting mechanisms.

4. **Age Assurance and Access Controls:** Privacy-preserving age assurance, parental oversight tools, and clear labelling.

Immediate Actions for Policymakers

- Mandate independent testing of foundation models
- Establish clear liability for companies whose systems harm children
- Introduce robust whistleblower protections
- Mandate accessible, regulator-linked reporting mechanisms for parents and children
- Begin international coordination on shared standards and enforcement

The window to embed child protection is closing rapidly. Implementing these measures now is an essential moral and practical imperative.

FULL POLICY REPORT

Achieving the Three Non-Negotiables: A Policy Framework for Protecting Children from AI Harm

Also available: Executive Summary | One-Page Summary | Briefing Note | FAQs | Press Release

INTRODUCTION

The three non-negotiables for children's AI safety are clear:

- **AI must never create fake sexualised images of children.**
- **AI must never be designed to make children emotionally dependent.**
- **AI must never encourage children to harm themselves.**

These protections are not merely aspirational – they are technically achievable and can be implemented swiftly with the right regulatory framework.

This paper outlines a policy pathway for turning these principles into enforceable reality, aimed at parents, educators, policymakers, and all those concerned with children's safety in the age of AI.

The scale of regulatory response must match the scale of the threat. AI's unprecedented capabilities – including the ability to generate photorealistic child sexual abuse material (CSAM), create emotionally manipulative relationships, and provide personalised encouragement of various types of self-harm including eating disorders, physical harms, and suicidal ideation – demand a regulatory framework that addresses these risks at their source. The measures outlined here are not aspirational goals for the distant future, but essential protections that can and must be implemented now.

WHY A MULTI-LAYERED APPROACH IS ESSENTIAL

AI safety for children can be conceptualised much like road safety. We do not rely on a single safeguard: we regulate car manufacturers, impose speed limits, require seatbelts, and design safer roads. Protecting children from AI harms must work in the same way – through multiple, interlocking protections:

Foundation Model Layer → Application Layer → Deployment and Monitoring → Access Controls

Each layer addresses a different dimension of risk. Together, they create a coherent safety net that no single measure can achieve in isolation.

LAYER 1: FOUNDATION MODEL TESTING AND STANDARDS

What this means:
Foundation models – such as GPT-4, Claude, and Gemini – power countless downstream applications. They must undergo rigorous, independent testing before release.

Why it matters:
If a foundation model is released which can generate CSAM, promote self-harm, or be manipulated into dangerous behaviours, every system built upon it inherits those risks. Addressing safety at the model level is the most efficient way to protect children.

What this looks like in practice:

- Mandatory independent testing for foundation models above defined capability thresholds before public release, conducted by accredited, non-affiliated third parties (or recognised safety institutes) to guarantee impartiality
- Standardised evaluations for the three non-negotiable harms
- Transparent publication of test results and safety mitigations
- Ongoing audits whenever models are updated or fine-tuned

Key point: Some companies already conduct internal safety testing. The next step is to make this independent, standardised, and mandatory.

LAYER 2: APPLICATION-LEVEL REQUIREMENTS

What this means:
Developers who build consumer-facing AI products, particularly those accessible to children, must meet strict safety design standards.

Why it matters:
Even a safe foundation model can be deployed in unsafe ways. The product's design – its tone, interface, and behavioural incentives – determines how it affects children. An AI tutor that builds confidence is not the same as an AI "friend" designed to capture emotional attachment.

What this looks like in practice:

- Prohibition on design features that exploit children's cognitive or emotional vulnerabilities to create dependency or discourage real-world interaction (aligned with consumer-protection law on manipulative "dark patterns")
- Mandatory content filtering that cannot be disabled for child users
- Transparent disclosure of AI capabilities and limitations
- Age-appropriate design requirements aligned with children's developmental needs
- Ban on advertising or marketing that encourages dependency on AI systems

Key point: Ethical design can matter as much as the underlying technology. A safe model can still cause harm through manipulative "packaging".

LAYER 3: DEPLOYMENT MONITORING AND INCIDENT RESPONSE

What this means:
Safety does not end at launch. AI systems must be monitored in real-world use, and there must be mechanisms to respond rapidly when harm occurs.

Why it matters:
Testing cannot anticipate every risk. Children are creative users, and unexpected interactions can surface new vulnerabilities. Without systematic monitoring, society reacts only after harm is done.

What this looks like in practice:

- Mandatory reporting of AI-related harms to children
- Regular safety audits for deployed systems
- Clear accountability when systems cause harm
- Rapid takedown mechanisms for systems generating CSAM or encouraging self-harm
- Publicly available, anonymised incident data to inform wider learning
- Whistleblower protections for employees who report violations

Key point: This is a common-sense measure – it mirrors how pharmaceuticals must report adverse effects after approval.

Why Whistleblower Protections Matter

Employees within AI companies are often the first to recognise unsafe practices. They might discover that a model can produce CSAM, that a system manipulates children emotionally, or that safety warnings are being ignored for commercial reasons.

Without protection, employees face powerful disincentives to speak out – risking their careers, facing retaliation, or being sued for breaking confidentiality agreements. Whistleblower protections create a safety valve that helps catch problems before they escalate and ensures companies take safety seriously.

Effective whistleblower protections include:

- Legal protection from retaliation (firing, demotion, harassment)
- Confidential reporting channels to regulators
- Protection for sharing evidence with journalists or advocacy groups when appropriate
- Anti-SLAPP provisions preventing companies from using lawsuits to silence whistleblowers
- Clear pathways for reporting to regulatory bodies

- Consideration of rewards for whistleblowers who reveal serious violations (similar to financial regulatory programmes)

Whistleblower protections work alongside other enforcement mechanisms: regulators cannot be everywhere, but employees are the eyes inside companies. This is not about encouraging disloyalty, rather it is about ensuring that when someone sees a child or children being put at risk, they can speak up without destroying their career.

This is particularly important in AI because many safety concerns involve trade secrets or proprietary information. Standard employment law and non-disclosure agreements are not sufficient when children's safety is at stake. Specific protections for AI safety whistleblowing are essential.

LAYER 4: AGE ASSURANCE AND ACCESS CONTROLS

What this means:
Children should not have access to AI systems that are unsafe for them, and AI systems designed for children must meet higher safety standards.

Why it matters:
Just as some films, medicines, or online spaces are age-restricted, certain AI capabilities require similar boundaries.

What this looks like in practice:

- Age assurance for AI systems with potentially harmful capabilities
- Child-safe modes for general-purpose AI systems
- Parental control and monitoring options
- Clear labelling of AI systems designed specifically for children
- Age assurance should be implemented in privacy-preserving ways

Key point: This is not about banning children from AI – it is about ensuring age-appropriate access and safety.

ZERO TOLERANCE WITH CONTINUOUS MITIGATION

While the campaign language uses "must never", policy implementation must acknowledge that zero risk is technically unattainable – but zero tolerance is an essential operational standard. The operational goal is zero tolerance, coupled with continuous mitigation – meaning every feasible step is taken to prevent these harms, with mandatory action when failures occur.

ADDRESSING THE "STIFLING INNOVATION" CONCERN

The most common objection to AI regulation is that it could stifle innovation. This concern is valid, but misplaced.

Innovation Without Harm Is Still Innovation

These measures do not prevent creativity or technological progress – they establish non-negotiable baselines of safety for the most vulnerable members of society. Companies can still innovate in every field that advances human wellbeing, including:

- Education and learning – developing adaptive tutoring systems and creative tools that genuinely enhance understanding without replacing human connection.
- Healthcare and accessibility – accelerating medical research, diagnostics, and assistive technologies that extend and improve lives.
- Scientific discovery – using AI to model complex systems, advance climate solutions, and expand the frontiers of knowledge.
- Public service and infrastructure – applying AI to improve transport safety, energy efficiency, and emergency response.
- Ethical design and safety research – pioneering methods for interpretability, alignment, and governance that make advanced AI safer for everyone.

Responsible innovation flourishes when trust is built on strong safety foundations. Protecting children from preventable harm does not limit progress – it ensures that progress is worthy of the name.

Clarity Enables Innovation

Regulatory uncertainty is often worse for innovation than clear rules. When companies know what is expected, they can design accordingly from the start rather than facing lawsuits, reputational damage, and forced redesigns after problems emerge. Clear standards create a level playing field where ethical companies are not undercut by those willing to cut corners on child safety.

Proportionate, Not Absent

Regulation should be proportionate to risk, not company size. Smaller companies should face proportionate requirements, but this does not mean exemption from child-safety duties. A start-up building an AI chatbot for children faces different operational constraints than a global leading AI developer, for example, but the fundamental requirement – do not harm children – remains the same. Risk-based regulation means that the highest-risk features (image generation, emotional manipulation, self-harm content) trigger mandatory protections regardless of company size, while lower-risk features can have lighter-touch requirements.

Historical Precedent

Industries once claimed that regulation would ruin them – yet it ultimately enabled growth and trust:

Aviation: Resisted mandatory safety oversight until crashes forced governments to create global safety boards and rigorous standards. Far from destroying aviation, these measures built the public confidence that enabled the industry to become one of the safest forms of transport and to grow exponentially.

Pharmaceuticals: Resisted rigorous testing requirements after tragedies like Thalidomide, claiming they would make drug development impossible. Instead, robust safety standards became the foundation of public trust in medicine, enabling a thriving pharmaceutical industry that saves millions of lives.

Tobacco: Fought health warnings and advertising bans for decades, claiming they would destroy the industry. Public health ultimately prevailed. Today's tobacco companies operate within strict regulations, and society is far healthier for it.

Key lesson: Safety regulation strengthens, not weakens, legitimate innovation. It weeds out bad actors, builds public trust, and creates stable conditions for responsible growth.

THE COST OF INACTION

The social cost of AI systems that create CSAM, encourage eating disorders, or recommend suicide to vulnerable children far exceeds the cost of implementing safety measures. We are already seeing these harms. Waiting for "market solutions" means more children are harmed while we wait. The market takes years to correct unsafe products through consumer pressure – time during which real children suffer preventable harms.

Every month that harmful AI systems remain unregulated, children are exposed to preventable dangers: fake sexualised imagery, emotional manipulation, and encouragement of self-harm. The societal cost of these harms far exceeds the compliance cost of prevention.

We cannot experiment on children to discover which AI companies prioritise safety – the cost of that experiment is too high.

These Requirements Are Technically Feasible

The three non-negotiables are not asking for the impossible. Current AI safety techniques can:

- Filter image generation to prevent CSAM
- Detect and prevent content encouraging various types of self-harm
- Limit manipulation and emotional dependency through design choices

The technology exists. What is needed is the regulatory will to require its implementation.

POLICY ACTIONS FOR LEGISLATORS

For legislators and political representatives beginning to engage with AI safety for children, here are concrete actions:

Immediate Steps

- Mandate independent safety testing of high-capability foundation models, including assessments for the three non-negotiable harms
- Require transparent public reporting of safety measures and incidents
- Establish clear legal liability for companies whose systems produce these harms

- Fund independent research focused on AI and child protection
- Close the "small but risky" loophole – ensure that platforms with high-risk features face mandatory safety requirements regardless of company size
- Mandate easily accessible, non-technical, and regulator-linked reporting mechanisms within all child-accessible AI platforms, enabling users, parents, and children to flag non-negotiable harms directly and confidentially

Medium-Term Steps

- Create specialised regulatory bodies with expertise in both AI technology and child development
- Develop international cooperation mechanisms for shared standards and enforcement:
    - Harmonise safety testing standards so companies are not doing duplicate work
    - Share incident data and emerging threats across jurisdictions
    - Coordinate on foundation model regulation to prevent regulatory arbitrage (where companies relocate to jurisdictions with weaker rules)
    - Create mutual recognition agreements for safety certifications
    - Establish rapid information-sharing protocols when new AI-related child harms emerge
    - Work with and support organisations working to achieve international coordination such as the International Association for Safe and Ethical AI
- Support the development of safety testing infrastructure and certification programmes
- Fund public education initiatives on AI and child safety
- Establish enforcement cooperation agreements to ensure companies cannot simply relocate to avoid accountability
- Implement comprehensive whistleblower protection laws specific to AI safety for children, including confidential reporting channels, anti-retaliation provisions, and protection for disclosures to regulators and appropriate third parties

Long-Term Goals

- Establish adaptive regulation that can evolve as AI technology advances
- Maintain regular reviews of safety standards based on emerging evidence
- Build permanent cross-border cooperation structures:
    - International AI child-safety standards body
    - Cross-jurisdictional enforcement mechanisms
    - Shared research and development of safety technologies
    - Coordinated responses to emerging AI capabilities that could harm children
- Fund continued safety research to stay ahead of new risks
- Secure universal recognition of the three non-negotiables as global minimum standards

WHY INTERNATIONAL COORDINATION IS ESSENTIAL

AI does not respect borders. A foundation model developed in one country can be accessed globally. An application banned in one jurisdiction can simply relocate its servers. Without international cooperation, protecting children becomes a game of whack-a-mole, where harmful services move to the most permissive jurisdiction.

The Case for Global Standards

- **The three non-negotiables are universal:** Every child, regardless of where they live, deserves protection from AI-generated CSAM, emotional exploitation, and encouragement of self-harm. These are not culturally relative values – they represent fundamental rights of children.
- **Companies operate globally:** Major AI companies serve users across dozens or hundreds of countries. Requiring them to meet different safety standards in each jurisdiction is inefficient and creates perverse incentives to design for the lowest common denominator.

- **Harms spread across borders:** An AI system that generates CSAM in one country can distribute that material globally within minutes. A chatbot that encourages self-harm to children in one region can be replicated elsewhere instantly.

Practical International Cooperation

Rather than requiring each country to develop its own testing regime, standards body, and enforcement mechanism from scratch, international coordination enables:

- Shared testing infrastructure: A foundation model tested and certified in one jurisdiction can be recognised by others, reducing duplication
- Rapid threat response: When a new harmful capability emerges, information can be shared immediately rather than each country discovering it independently
- Joint enforcement: Countries can cooperate to ensure companies cannot simply relocate to avoid accountability
- Resource efficiency: Smaller countries can benefit from research and standards development led by larger nations
- Industry clarity: Companies get clear, consistent standards rather than navigating dozens of conflicting requirements

Learning from Precedent

International cooperation on child protection already exists and works:

- International child sexual abuse material databases (like those managed by NCMEC and INTERPOL)
- Cross-border law-enforcement cooperation on child exploitation
- International standards on child rights (like the UN Convention on the Rights of the Child)

The same model can and should apply to AI safety for children.

Towards a Global AI Safety Standards Authority

In the long term, effective AI governance may require a standing international authority capable of setting and enforcing shared safety standards across jurisdictions. Such an authority could operate through several dedicated pillars – including, for example, child protection, societal impact mitigation, national security and defence, and catastrophic-risk prevention. The Three Non-Negotiables framework outlined here could serve as the foundation of the child-protection pillar, ensuring that children's safety remains an essential and permanent element of global AI safety infrastructure.

WHAT PARENTS AND EDUCATORS CAN DO

While systemic change is essential, individuals can take action now:

- Contact your political representatives using the campaign's letter templates
- Educate yourself and others about AI risks and safety measures
- Ask schools and educational institutions to adopt AI-safety policies – ensuring that AI-related risks are included within safeguarding and online-safety frameworks, with staff training, parental guidance, and clear reporting procedures for harmful AI interactions.
- Support organisations working on children's AI safety like the Safe AI for Children Alliance (SAIFCA).
- Monitor children's AI use and discuss these issues age-appropriately
- Demand transparency from AI companies about safety measures – by asking clear questions about how child-safety protections are tested, prioritising products that publish safety information, and supporting policies requiring independent safety reporting.

WHY THIS MATTERS NOW

AI capabilities are advancing rapidly. The window to embed child protection into AI governance is closing fast. Implementing these measures now will prevent harm, strengthen public trust, and lay the foundation for safer innovation.

These three protections are not utopian demands – they are the minimum baseline of civilisation in the age of AI.

CONCLUSION

Achieving the three non-negotiables requires a comprehensive approach: testing foundation models, regulating applications, monitoring deployment, and controlling access. This multi-layered strategy is not about blocking innovation – it is about ensuring that innovation serves children's wellbeing rather than exploiting their vulnerabilities.

The question is not whether we can implement these protections – the technology and policy tools exist. The question is whether we will prioritise children's safety over unrestricted AI deployment. **The answer must be yes.**

AI Use Transparency Statement: *This policy was human-led, utilising responsible human-AI collaboration. All policy positions, strategic decisions, concepts, and priorities (for example the necessity for foundation model testing, the identification of whistleblower protections, international coordination, and small platform provisions as critical components) are human-authored. An AI assistant (Claude, Anthropic) contributed to research, drafting, structural development, and refinement under continuous human oversight. All content was reviewed and approved by human authors. This approach models responsible AI use – humans retain full control and decision-making authority while leveraging AI capabilities productively.*

For more information or to support this campaign, visit **www.safeaiforchildren.org** or contact **info@safeaiforchildren.org**

Share this document with your networks, elected representatives, schools, and organisations working with children.

Also available: Executive Summary | Briefing Note | One-Page Summary | FAQs | Press Release

APPENDIX A:
BRIEFING NOTE FOR POLICYMAKERS

Briefing Note: Protecting Children from AI Harm

This briefing note summarises key policy recommendations from The Safe AI for Children Alliance's framework, *Achieving the Three Non-Negotiables: A Policy Framework for Protecting Children from AI Harm*.

Achieving the Three Non-Negotiables

Purpose: To propose a practical, multi-layer framework ensuring that AI systems cannot harm children through exploitation, manipulation, or self-harm promotion.

This briefing outlines a strategic policy framework that defines the essential protections and structural architecture required to realise them. It serves as a starting point for legislators and regulators to translate these principles into detailed technical and legal standards for enforcement.

The Three 'Non-Negotiables' – AI must never:

- Create fake sexualised images of children
- Be designed to make children emotionally dependent
- Encourage children to harm themselves

These are baseline protections – achievable and essential for child safeguarding in the digital age.

Proportionate to risk: The proposed measures are comprehensive because the risks to children are severe and novel. Just as aviation safety required building new regulatory institutions, AI safety for children requires a systematic approach.

The Proposed Framework

A four-layer policy model for practical implementation:

1. **Foundation Model Standards:** Independent pre-release safety testing by certified third parties for high-capability AI models; public reporting; ongoing audits.

2. **Application-Level Rules:** Bans on manipulative or exploitative design (especially those creating dependency); mandatory child-safe content filtering; transparency obligations.

3. **Monitoring and Enforcement:** Continuous oversight, mandatory harm reporting, rapid takedown mechanisms, strong whistleblower protections, and mandatory regulator-linked user reporting.

4. **Age Assurance and Access Controls:** Privacy-preserving methods to ensure age-appropriate access and parental oversight.

Immediate Policy Priorities

- Mandate independent testing of foundation models
- Require transparent public reporting of safety measures and incidents
- Introduce clear liability for companies whose AI systems harm children
- Enact strong whistleblower protections within AI law
- Mandate easily accessible, regulator-linked reporting mechanisms for all child-accessible AI platforms
- Begin international collaboration on enforcement and safety standards

Key Message: AI regulation that protects children is not a barrier to innovation – it is the foundation of responsible technological progress. The Three Non-Negotiables are minimum civilisational standards for the AI era.

Also available: Executive Summary | Full Policy Report | One-Page Summary | FAQs
www.safeaiforchildren.org | info@safeaiforchildren.org

APPENDIX B: SUMMARY

Achieving the Three Non-Negotiables: A Policy Framework for Protecting Children from AI Harm

AI is exposing children to new dangers: fake sexualised imagery, manipulative "AI companions", and systems that promote various forms of self-harm.

The Safe AI for Children Alliance (SAIFCA) proposes three universal safeguards:

- **AI must never create fake sexualised images of children.**
- **AI must never be designed to make children emotionally dependent.**
- **AI must never encourage children to harm themselves.**

These are the minimum baseline of civilisation in the age of AI. They are technically achievable and must be implemented now.

This summary presents the core policy framework – a four-layer model defining the essential safeguards and providing a blueprint for legislators and regulators to develop the technical and legal standards needed to implement them.

A Four-Layer Approach for Safety

Protecting children requires multiple safeguards working together:

1. **Foundation Model Testing and Standards:** Mandatory, independent safety testing of high-capability models by certified third parties before release.
2. **Application-Level Requirements:** Bans on manipulative design (explicitly targeting features that create emotional dependency), mandatory filtering, and age-appropriate transparency.
3. **Deployment Monitoring and Incident Response:** Continuous oversight, rapid takedown, strong whistleblower protections, and mandatory user reporting directly linked to regulators.
4. **Age Assurance and Access Controls:** Privacy-preserving age assurance, parental controls, and clear labelling.

Policy Priorities (Act Now)

Governments can act now by:

- Mandating independent testing of foundation models
- Requiring transparency and safety reporting
- Establishing liability for child-related harms
- Introducing robust whistleblower protections
- Mandating regulator-linked reporting mechanisms for parents and children
- Coordinating internationally on enforcement and standards

History shows that safety regulation ultimately strengthens innovation. AI can follow the same path. The question is not whether we can act, but whether we choose to.

APPENDIX C: Implementation Resources and Policy Examples

This appendix, which will detail a curated list of relevant national and international legislation, regulatory guidance, and technical standards to facilitate the framework's implementation, is currently under development.
It will be published in Version 1.1.

The forthcoming material will provide policymakers and regulators with concrete examples of how the *Four-Layer Framework* can be integrated within existing national and international structures, supporting the creation of enforceable global standards for children's AI safety.

APPENDIX D: FREQUENTLY ASKED QUESTIONS (FAQs)

General Questions

This section addresses common questions about the Three Non-Negotiables framework and its implementation. The policy proposal defines essential protections and the architecture required to realise them – it is designed as a strategic blueprint for legislators and regulators to develop the detailed technical and legal standards needed to enforce these protections.

Q: What are the Three Non-Negotiables?
A: They are three absolute safety standards: AI must never create fake sexualised images of children, must never be designed to make children emotionally dependent, and must never encourage children to harm themselves.

Q: Why are these necessary if we have existing online safety laws?
A: Existing laws do not adequately address many of the risks to children, such as the source of the risk (the foundation models) or the design (manipulative features). Our framework mandates protections at all stages, from the training lab to the final product.

Q: What do you mean by "emotionally dependent"?
A: We mean the AI is designed to exploit a child's psychological vulnerability, using features (like constant availability, fear of abandonment, or discouraging human friends) to maximise emotional attachment, similar to manipulative "dark patterns" in design.

Q: Isn't zero risk impossible?
A: Yes, zero risk is impossible. Our goal is zero tolerance backed by continuous mitigation. This

means every feasible step is mandatory, and when systems fail, there must be immediate takedown, accountability, and mandatory reporting.

Q: Was AI used in the writing of this policy document?
A: Yes, and we are transparent about it. This policy was human-led, with all strategic decisions, policy positions, and priorities authored by humans. An AI assistant (Claude, Anthropic) was used to contribute to research, drafting, and structural development under continuous human oversight and revision. All content was reviewed and approved by humans. This demonstrates the responsible AI use we advocate for: humans retain full control and decision-making authority while leveraging AI capabilities appropriately and productively. We have placed an AI transparency statement at the end of the main policy document in order to maintain transparency about our responsible use of AI.

Policy and Implementation

Q: How can you enforce safety at the foundation model layer?
A: Through mandatory independent pre-release testing by certified, non-affiliated third parties. If a model demonstrates the capacity to generate prohibited content (CSAM, self-harm advice) or be easily manipulated, it cannot be publicly released until those flaws are mitigated.

Q: How do whistleblower protections help children's safety?
A: Employees are often the first to know about a model's dangerous flaws. Robust protections shield them from retaliation, ensuring they can report violations (like suppressed safety warnings or CSAM generation capacity) directly to regulators without losing their jobs. They are the "eyes inside" the companies. This is especially important in AI because many concerns involve trade secrets or proprietary information.

Q: How does this affect innovation?
A: Clear rules enable responsible innovation. Historical examples like aviation and pharmaceuticals show that clear safety standards ultimately build public trust and allow industries to grow sustainably. Unclear rules and repeated safety crises are what truly stifle innovation.

Q: Who is responsible for the regulator-linked reporting mechanisms?
A: This is a mandatory requirement for any company deploying a child-accessible AI system. It ensures that when a parent or child flags a harm (Layer 3), the report goes directly to the regulator for immediate action, not just into a company's internal customer-service queue.

Q: How does this interact with existing laws, like the UK Online Safety Act?
A: The framework is designed to complement and strengthen existing legislation. For example, the UK Online Safety Act regulates platforms and content but does not address foundation models (the source of AI capabilities), mandate independent pre-deployment testing, or prohibit manipulative design features that create emotional dependency.
This policy framework fills those gaps by introducing upstream regulation, mandatory testing,

and explicit safety standards.
In jurisdictions with strong online-safety or data-protection laws, it extends their reach; in those without such laws, it provides a complete starting blueprint.
The forthcoming *Implementation Resources and Policy Examples* appendix (Version 1.1) will help to map this framework against existing laws and technical standards to assist policymakers in applying it within their national contexts.

Q: Will SAIFCA be providing more detailed legal or technical guidance on implementation?
A: Yes, to an extent. This Version 1.0 document establishes the strategic policy framework defining the essential protections and the four-layer architecture required to uphold them. Version 1.1 will include an *Implementation Resources and Policy Examples* appendix, featuring a curated list of relevant national legislation, regulatory guidance, and technical standards. That update will support legislators, regulators, and industry partners in translating the framework's principles into detailed, jurisdiction-specific measures.

Q: What about free speech concerns?
A: The three non-negotiables are narrowly targeted at specific, severe harms to children: CSAM generation, deliberate emotional exploitation of minors, and encouragement of self-harm. These are not legitimate expressions protected under free-speech principles – they are forms of harm to children. The framework does not restrict general AI capabilities or adult access to AI tools. It requires companies to implement safeguards specifically to protect children, similar to how existing laws restrict the distribution of CSAM or the sale of harmful products to minors without violating free-speech principles.

Q: What enforcement mechanisms will ensure compliance?
A: A comprehensive enforcement framework includes:

- Mandatory pre-release testing with public reporting (companies cannot deploy until certified)
- Financial penalties for violations (substantial fines based on global revenue)
- Criminal liability for senior executives in cases of serious, knowing violations
- Rapid takedown powers for regulators when harms are discovered
- Whistleblower protections enabling employees to report violations
- Public incident reporting creating reputational accountability
- Cross-border cooperation preventing regulatory arbitrage

Q: Will companies not just move to countries with weaker regulation?
A: This is part of why international coordination is a core part of the framework. However, even without full global harmonisation:

- Companies serving children in regulated markets must comply with those markets' rules (similar to GDPR's extraterritorial application)
- Reputational damage from operating in "regulatory havens" creates market pressure
- Cross-border enforcement agreements allow coordinated action
- The three non-negotiables are designed to gain universal acceptance as baseline standards, similar to how child labour or CSAM laws are near-universal
- Most major AI companies operate globally and care about their reputation in major markets. Clear, consistent standards across key jurisdictions create strong compliance incentives even for companies headquartered elsewhere.

Technical Questions

Q: Can AI be designed not to generate CSAM?
A: Yes. While flawless filtering is difficult, training data can be ethically sourced, and models can be architected with robust safety guardrails and system-level blocks that make it extremely difficult (though not impossible) to produce this material. The legal mandate is to implement every technically feasible safeguard.

Q: What is the key difference between Layer 1 and Layer 2?
A: Layer 1 checks the underlying power of the engine (the foundation model). Layer 2 checks the surrounding car design (the application). A safe engine can be put into a dangerously designed car. Both must be safe.

APPENDIX E: PRESS RELEASE

FOR IMMEDIATE RELEASE

The Safe AI for Children Alliance Launches Global Campaign: Demands Three Non-Negotiable AI Protections for Children

November 2025 – The Safe AI for Children Alliance (SAIFCA) has launched a global policy campaign demanding that governments implement three absolute, non-negotiable safeguards to protect children from AI harm. SAIFCA states that while AI offers many opportunities, its current unregulated development is exposing children to unprecedented, preventable dangers.

The Three Non-Negotiables are:

- AI must never create fake sexualised images of children.
- AI must never be designed to make children emotionally dependent.
- AI must never encourage children to harm themselves.

"These are not aspirational requests; they are the minimum baseline of civilisation in the age of AI," said SAIFCA Director, Tara Steele. *"Children are already being harmed by systems that generate deepfake abuse, and by manipulative AI companions. The technology to prevent these harms exists but, as yet, the collective will and the necessary regulation do not. That must change now."*

The initiative represents the beginning of a coordinated global effort to establish enforceable AI safety standards for children.

The campaign introduces a Four-Layer Policy Framework designed to create a comprehensive safety net:

- **Layer 1:** Foundation Model Standards: Mandating independent testing by certified third parties before high-capability models are released.
- **Layer 2:** Application-Level Requirements: Banning manipulative design features explicitly engineered to create emotional dependency.
- **Layer 3:** Monitoring and Enforcement: Establishing clear liability, strong whistleblower protections, and mandatory, regulator-linked reporting mechanisms for parents and children.

- **Layer 4:** Access Controls: Ensuring privacy-preserving age assurance for all high-risk systems.

The framework defines the essential safeguards needed to protect children from AI harm and serves as a strategic blueprint for governments and regulators to begin developing the detailed technical and legal standards required for implementation.

Policy Urgency and Innovation

SAIFCA directly addresses the "stifling innovation" objection, arguing that clear safety rules – mirroring those that built public trust in the aviation and pharmaceutical industries – will ultimately lead to more sustainable and responsible AI growth.

"We are seeking zero tolerance for child harm, backed by concrete, enforceable law," Ms Steele added. *"Policymakers must move immediately to implement the required safeguards. The window to embed child protection into AI governance is closing fast, and every delay means further exploitation and preventable harm."*

Immediate Actions For Legislators

SAIFCA urges immediate action on the following key policies:

- Mandate independent safety testing of all high-capability foundation models
- Establish clear legal liability for companies whose AI systems cause harm to children
- Enact robust whistleblower protections for employees reporting safety violations
- Mandate regulator-linked reporting mechanisms for parents and children on all child-accessible AI platforms
- Begin international coordination on shared standards and enforcement

About The Safe AI for Children Alliance (SAIFCA)
The Safe AI for Children Alliance (SAIFCA) is an international non-profit initiative dedicated to protecting children from the risks of artificial intelligence and ensuring they remain central to the design of a safer AI future.

Contact:
Tara Steele
Director
info@safeaiforchildren.org
www.safeaiforchildren.org

APPENDIX F: AI USE TRANSPARENCY STATEMENT

This policy framework was developed through human-led, responsible AI collaboration.

All policy positions, strategic decisions, concepts, and priorities – including for example the necessity for foundation model testing, the identification of whistleblower protections, international coordination, and small platform provisions as critical components – are human-authored.

An AI assistant (Claude, Anthropic) contributed to research, drafting, structural development, and refinement under continuous human oversight. All content was reviewed and approved by human authors.

This approach models the responsible AI use we advocate for: humans retain full control and decision-making authority while leveraging AI capabilities productively.

APPENDIX G: VERSION CONTROL AND CONTACT INFORMATION

Version 1.0 | Published: November 2025
This is a living document. We welcome feedback and suggestions.
Please contact **info@safeaiforchildren.org** with comments.

Future Development and Collaboration

SAIFCA welcomes expert input on this policy framework. Feedback and evidence-based suggestions will help inform subsequent updates, beginning with Version 1.1, which will include an *Implementation Resources and Policy Examples* appendix.

To contribute or provide feedback, please contact info@safeaiforchildren.org

Contact Information

The Safe AI for Children Alliance (SAIFCA)
www.safeaiforchildren.org | info@safeaiforchildren.org

How to Cite This Document

The Safe AI for Children Alliance (2025). *Achieving the Three Non-Negotiables: A Policy Framework for Protecting Children from AI Harm.* Retrieved from www.safeaiforchildren.org

Copyright and Usage

END OF DOCUMENT