BRIEFING NOTE: PROTECTING CHILDREN FROM AI HARM

The Three Non-Negotiables Policy Framework

THE SAFE AI FOR CHILDREN ALLIANCE (SAIFCA)

VERSION NUMBER 1.0

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:

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

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

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

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.

End of Briefing Note