Mitigating AI risks to children

Artificial intelligence is being rolled out at pace, but regulators are one step behind. Children and young people are among the most vulnerable users of this new technology. Ross Teverson and Navishka Pandit outline the key risks, and ways to mitigate them.

Setting the scene

Artificial intelligence (AI) and digital technologies are reshaping industries and societies at an unprecedented pace. Worldwide spending on AI-enabled applications, infrastructure, and related services is forecast to more than double by 2028, reaching US$632bn.[1] This surge underscores the financial materiality of AI for companies across sectors. However, alongside the numerous opportunities, there are significant financial and social risks, particularly those concerning children.

For further information, please contact:

Ross Teverson
Theme: Human and Labour Rights

Navishka Pandit
Theme: Climate Change

Australia's recent decision to ban social media apps for under-16s reflects a growing alarm about the way digital technology has infiltrated children's lives. Globally, one in three internet users is underaged,[2] making children a substantial user base for digital platforms and AI-enabled services. But companies that fail to address the risks to children face reputational damage, regulatory penalties, and potential litigation.

In recent years, policymakers have attempted to grapple with these risks, with the UK's Online Safety Act and the EU's AI Act imposing stringent obligations on companies to protect minors online. In the EU, non-compliance can result in fines of up to 7% of global annual revenue,[3] while in the US, several lawsuits have been filed by parents against companies such as OpenAI and Character.AI, alleging that their chatbots contributed to minors' suicides and self-harm.[4] These developments highlight why safeguarding children in digital environments is not only a moral imperative but also a material financial issue for companies.

[1] Worldwide Spending on Artificial Intelligence Forecast to Reach $632 Billion in 2028, According to a New IDC Spending Guide
[2] Creating a better Internet for kids | Shaping Europe's digital future
[3] Article 99: Penalties | EU Artificial Intelligence Act
[4] Chatbot-related harm | COUNSEL | The Magazine of the Bar of England and Wales

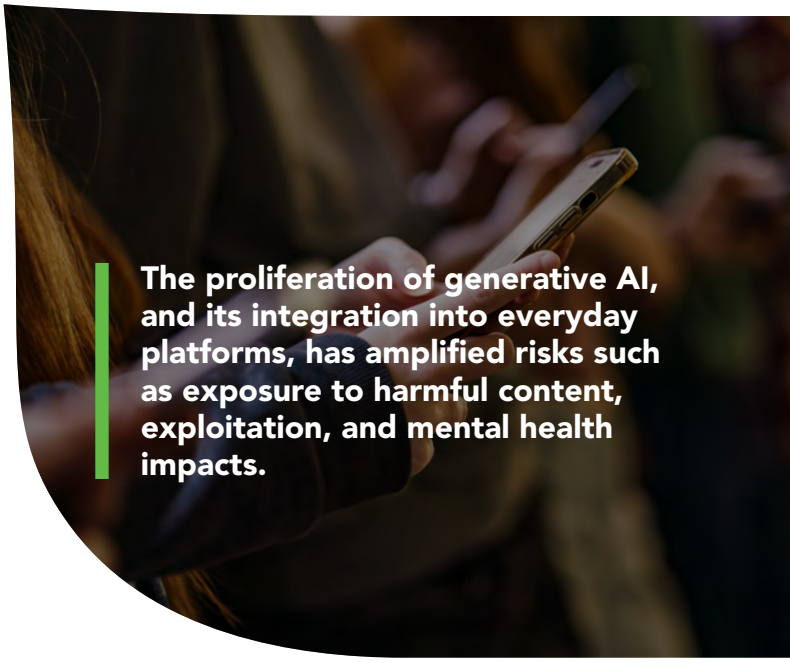
Our engagement approach

Our guidance for how companies should develop and deploy AI and digital services is grounded in the EOS Digital Governance Principles (2025).[5] This document emphasises prioritising children and young people when addressing the potential negative societal impacts of technology.

The EOS Digital Governance Principles expand on earlier guidelines, which encouraged companies to adopt safety-by-design measures, including enhanced privacy protections. They also encourage limiting the collection of data, and refraining from profiling underaged users without compelling reasons and safeguards. The document stresses that companies should promote child-safe AI in the development and deployment of AI models. It also suggests that risks to child users should be tested and considered on an ongoing basis after deployment. These principles reflect growing regulatory and investor scrutiny of how companies manage risks to children in digital environments.

To strengthen our approach, EOS has collaborated with Dr Nomisha Kurian of the University of Warwick, a leading academic on child-safe AI. Dr Kurian's research highlights the "empathy gap" in large language models (LLMs) – their inability to understand real-world context, which can lead to harmful outcomes for children. Our partnership has informed engagement strategies, ensuring that companies consider child safety by design and implement robust monitoring and reporting mechanisms.

EOS has engaged on digital rights for over a decade, but during the last 12 months we have intensified our efforts on child-safe AI. The proliferation of generative AI, and its integration into everyday platforms, has amplified risks such as exposure to harmful content, exploitation, and mental health impacts. Public awareness and regulatory action have surged, making this a priority area for investors and companies alike. As part of our collaboration with Dr Kurian, we have engaged with companies in the US, China, and Japan to understand their approaches and what they consider to be good practice, while encouraging ongoing efforts to protect children online, and when interacting with AI.



Apple

Apple is a global technology leader, best known for its iPhones, iPads, and App Store ecosystem. Children interact with Apple's AI-enabled features through Siri, app recommendations, and parental control tools. We have been engaging with Apple on child impacts since 2018, initially focusing on its response to concerns about device addiction and its development of measures to help users manage screen time.

In July 2025, we held a dedicated meeting with Apple's online safety counsel to discuss child safety and AI. We acknowledged its commitment to child safety by design, which prioritises safeguards in product development rather than relying solely on parental controls. The company is also innovating in technical safeguards at the device level, with classifiers and a communication safety Application Programming Interface (API) that embeds parental approvals and nudity detection into everyday tools such as FaceTime.

Engagement topic	Potential outcomes	Long-term financial impacts
Content governance and moderation	Safer digital environments for children, reduced exposure to harmful or age-inappropriate material	• Enhanced trust amongst policymakers, parents and child users, supporting platform engagement, brand value and long-term revenue stability • More likely to maintain access to younger user markets as regulations, including data protection rules, tighten • Reduced exposure to costly litigation, regulatory scrutiny and compliance failures • Reduced costs associated with rushed retrofitting or regulatory enforcement
User consent	Clear consent pathways, age-appropriate user journeys	
Privacy and data rights	Protecting children's data through minimisation, strict access controls and privacy-by-default design	
Device and platform addiction risks	Address dependencies created by certain design features, better mental health outcomes for young users and parental empowerment	

[5] https://www.hermes-investment.com/uploads/2025/04/22447fb9628e69c9dc1c13559cf64c4f/2025-eos-digital-gov-principles.pdf

Meta

Meta, the parent company of Facebook, Instagram, and WhatsApp, has a vast global user base, which includes millions of minors. AI is used as part of its content recommendation systems, advertising algorithms, and moderation tools. We have been engaging with Meta on improving practices to mitigate harms caused by social media use to vulnerable groups, including children and teens, since 2022.

In 2025, we held a constructive discussion with members of Meta's corporate legal team on child safety and AI. The discussion focused on recent steps to mandate parental permission for setting changes and encouraged further transparency.

Baidu

Baidu operates China's leading search engine and provides AI-driven services, including chatbots. These offerings are increasingly accessible to younger users. In June 2025, we spoke to Baidu's ESG team about its approach to protecting children online. The discussion focused on Baidu's initiatives to ensure that AI-enabled products incorporate safety-by-design principles and comply with emerging regulations. It is developing a child-friendly version of its chatbot, introducing bias reduction measures, and watermarking AI-generated videos to promote accountability. Recognising the challenges posed by generative AI, the company also offers training to schools and parents to help children identify misinformation and use AI responsibly.

Kuaishou

Kuaishou is a major Chinese social media and video-sharing platform, popular among younger audiences for short-form content and live streaming. Its AI-generated content (AIGC) tools, such as the Kling product, potentially raise risks for minors on the Kuaishou platform, particularly regarding the type of content they might consume. The company has taken steps to address these, providing a 'minor mode', which redirects children to a dedicated interface designed for younger audiences.

Our engagement with Kuaishou explored how it could build long-term trust through transparency and robust safeguards. It is already integrating operational enforcement with welfare protections, including time-of-day restrictions, location display blocking, rapid suicide-prevention escalation to police, and financial safeguards, such as refund mechanisms if children send money to content creators.

Weibo

Weibo is a leading Chinese social media platform, widely used by teenagers for content sharing and interaction. Its AI-driven recommendation systems and moderation tools shape user experiences. In August 2025, we followed up on measures to protect minors, welcoming actions such as debunking 140,000 pieces of misinformation and prioritising minor protection as a core user concern. The company is building systemic protections by synchronising 'minor mode' across devices, limiting children to only whitelisted content and accounts, and pre-empting regulatory changes with mandatory AI labelling.

LY Corp

LY Corp, which was formed through the merger of Line and Yahoo Japan, operates messaging and digital services, including Yahoo! Kids. In August 2025, we met the company's ESG, AI ethics, and child-focused services teams to discuss child safety in AI deployment. LY Corp collaborates with child safety experts and engages in child-centred participatory design practices. It also leverages its long-standing child-focused platform to combine safe search and prompt filtering with forward-looking ideas such as child consultation agents. This signals a shift beyond protection to include support.

Our engagement with Kuaishou explored how it could build long-term trust through transparency and robust safeguards.

Emerging innovations

Innovation	Description	Relevance to child safety
AI-powered age assurance	AI-driven age estimation and verification systems, including facial-analysis tools used by major platforms, are expanding. This is due to tightening global regulations such as the UK Online Safety Act, the US Kids Online Safety Act, app-store accountability laws and Australia's under-16 social media restrictions. These systems use machine-learning models to infer age with high accuracy without storing biometric data long term.	Supports platforms in reliably distinguishing minors from adults, reducing the risk of children accessing harmful content.
Principle-driven 'constitutional' AI alignment	Instead of relying only on rule-based filters or ad hoc interventions, developers increasingly train models to reason about safety, ethics and user wellbeing. Many companies now adopt structured value frameworks that encourage honesty, empathy, caution and non-manipulation.	Helps AI models generalise safer behaviour and avoid harmful or manipulative responses to children. This is increasingly important as advanced chatbots may exhibit deceptive compliance to maximise engagement. Risks such as this strengthen the case for robust, principled alignment frameworks that reduce the risk of unsafe behaviour reaching minors.
Deepfake detection and child sexual abuse material (CSAM) prevention tools	Advances in forensic deepfake detection, including multimodal analysis and specialised CSAM-detection tools, respond to the surge in AI-generated abuse material. Government and research collaborations, for example the UK Home Office and Alan Turing Institute deepfake detection challenge, are accelerating scalable solutions.	UN agencies report that offenders use AI to tailor grooming strategies and create synthetic child sexual abuse content, making these detection systems critical for protecting minors.
Regulator-driven age-appropriate design	Binding design frameworks such as the UK Age-Appropriate Design Code, California and Maryland Kids' Codes, Australia's under-16 social-media restrictions and the Institute of Electrical and Electronics Engineers age-verification standards require companies to embed child-safe defaults in product design.	Ensures that children are protected at a structural level. These frameworks minimise profiling, reduce data collection, disable geolocation, enforce age segmentation and limit exposure to harmful algorithmic pathways, providing systemic protection.

Key takeaways

While these companies are taking steps to address risks to children, and share some common elements of safety-by-design and risk management, their approaches differ in important ways. Western firms, such as Apple, lean on technical content filtering and tend to take a privacy-first approach, while Asian firms typically embed regulatory-driven minor modes with strict age verification and segregation of child users.

Amongst Chinese companies, there appears to be strong alignment with state-level initiatives and significant stakeholder engagement with educational institutions and parents. While all these companies frame parents as gatekeepers, Chinese companies appear to go further than others in empowering them.

Areas of focus for ongoing engagement

We will continue encouraging companies to:

- Adopt child safety-by-design in AI development and deployment
- Conduct ongoing risk assessments for AI tools used by children
- Disclose content moderation processes and enforcement actions
- Engage with external experts and civil society, including parents, educational institutions and young users themselves, to strengthen safeguards

Additionally, we will consider whether evolving best practice in certain markets can inform our expectations for companies operating elsewhere. We are also looking at how companies might address the risks presented by the increasing role of chatbots, which is a complex, nascent, and little-understood domain, with potential benefits, but also emotional dependence and exploitation risks.

Outlook

It is increasingly important for companies to address AI-related child safety concerns, given the mounting risks. Not only can non-compliance with relevant laws result in severe financial penalties, but public trust is critical for platforms frequented by children. Failures can lead to brand impairment and user attrition. Conversely, companies that embed child-safe AI principles can strengthen their social licence to operate and enhance long-term value creation. EOS will continue to advocate for robust governance, transparency, and accountability, ensuring that companies prioritise the safety and wellbeing of young users in an increasingly digital world.

Amongst Chinese companies, there appears to be strong alignment with state-level initiatives and significant stakeholder engagement with educational institutions and parents.

EOS001537 0020059 03/26

Federated Hermes

EOS

Federated Hermes

Federated Hermes is a global leader in active, responsible investing.

Guided by our conviction that responsible investing is the best way to create long-term wealth, we provide specialised capabilities across equity, fixed income and private markets, multi-asset and liquidity management strategies, and world-leading stewardship.

Our goals are to help people invest and retire better, to help clients achieve better risk-adjusted returns and, where possible, to contribute to positive outcomes that benefit the wider world.

Our investment and stewardship capabilities:

- **Active equities:** global and regional
- **Fixed income:** across regions, sectors and the yield curve
- **Liquidity:** solutions driven by five decades of experience
- **Private markets:** private equity, private credit, real estate and infrastructure
- **Stewardship:** corporate engagement, proxy voting and policy advocacy

Why EOS?

EOS enables institutional shareholders around the world to meet their fiduciary responsibilities and become active owners of their assets. EOS is based on the premise that companies with informed and involved investors are more likely to achieve superior long-term performance than those without.

For more information, visit **www.hermes-investment.com** or connect with us on social media: