Charter of Engagement for Safer AI for Children

Coalition for Children's Rights and Protection in the Age of AI

Preamble

Generative AI is entering environments that matter for children's development, including education, play, communication, information-seeking, creative work, emotional support, and family life. Over time, this technology may shape, how young people develop, learn, relate, participate and contribute in society.

AI could offer significant opportunities in areas such as learning, creativity, accessibility, communication and support, as well as in preventing, detecting and responding to child sexual abuse and exploitation. These benefits need to be considered alongside the fact that these systems can be fluent, profoundly personalized, adaptive, multimodal and capable of socially or emotionally responsive interaction, with the risk of generating negative impacts on children. The goal is to realize the benefits of generative AI while building the evidence, shared understanding and safeguards needed to mitigate risks on social and cognitive development, before patterns of use become harder to change.

The risks and opportunities of generative AI should be assessed through a developmentally differentiated lens, where uses, features and interaction patterns, are considered across age groups, developmental contexts and levels of use.

The evidence base on repeated child and adolescent interaction with generative AI is still developing, and mature longitudinal research is not yet available. However, adoption at scale is already underway, and product and policy decisions are already being made.

In this context, existing developmental science, adjacent evidence, early empirical findings, product knowledge and expert judgment provide the strongest available basis for action, but that evidence needs to be organized around the priorities stakeholders are already facing. While safeguards should be implemented now, in proportion to the identified risks, they need to be frequently refined and updated, as the evidence gaps are filling. The Coalition for Children's Rights and Protection in the Age of AI, led by France, Spain, Kenya and involving ten United Nations agencies, brings together governments, researchers, civil society and industry to close that gap. The purpose of this Charter is to strengthen independent evidence on risks and opportunities of AI for child development and bring that evidence into product design, evaluation and public policy.

This Charter is the contribution of its industry partners to a broader multi-stakeholder dialogue involving governments, academia, researchers, civil society and other stakeholders committed to advancing safe and age-appropriate AI for children. It builds on a growing body of political and normative commitments of its members [1].

[1] The G7 Evian Call for a Safer Digital Space for Minors, endorsed by G7 Leaders together with India, Brazil, Kenya, the Republic of Korea, Egypt, Australia and Spain ; the G7 Common Principles for a safer and more protective digital space for minors endorsed by the G7 Digital and Technology Ministers in June 2026; the principles of the UN Convention on the Rights of the Child ; the Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse (2020), the Joint Statement on Artificial Intelligence and the Rights of the Child, endorsed by more than 50 organisations under the leadership of ITU and the Committee on the Rights of the Child, principles set out in United Nations General Assembly Resolution A/78/L.49, "Seizing

Shared principles

This coalition is guided by the following principles, intended to inform the design, implementation, operation, and risk mitigation of AI systems that children may access:

Safety by design and by default: Take concrete measures to ensure a high level of safety, privacy and security for children by design and by default, applying relevant mitigations that protect children's safety and well-being. Such may include implementing age-appropriate time-management and agency-enhancing tools to guard against excessive and compulsive uses or services that foster emotional attachment.

AI that Prioritizes the Best Interests of children: Uphold children's rights and foster AI solutions developed by using child-centered approaches that prioritize their developmental needs and preserve offline times. This includes transparent practices, appropriate parental control to align AI systems with children's rights, needs, and well-being. This includes supporting international research on anthropomorphism, over-reliance and emotional dependency.

Ethical, Equitable, and Inclusive AI: Implement AI that strives to benefit all children regardless of background, ability, or needs, bridging the digital gap. AI systems must consistently seek to promote inclusivity and enable equal accessibility.

Age-Appropriate AI by Design: Create and use AI solutions aligned with children's cognitive and socio-emotional stages accounting for differences in maturity, vulnerability and development, as well as rights to information and participation.

Family-Inclusive Design: Provide parents and guardians with effective, accessible and user-friendly tools, and which are interoperable where technically feasible, together with supportive resources so they can confidently manage children's AI experiences. This also includes considering children's perspectives, needs and experiences to inform how companies can approach this space.

Meaningful Child Participation: Support the meaningful, age-appropriate and ethical participation of children and young people in research, ensuring that their perspectives, experiences and needs to help inform the development of recommendations, best practices and approaches to AI systems affecting them.

Privacy for Children: Protect children's data privacy with robust safeguards and data minimization, explain in age-appropriate ways how data is collected, used, and protected and transparency with children's unique interests in mind.

Effective and privacy-preserving age assurance: Implement robust, reliable, accurate, risk-based and privacy preserving, non-discriminatory and interoperable age-assurance mechanisms that do not solely rely on self-declaration, and support the independent evaluation of the accuracy, privacy and fairness of age assurance solutions.

---

the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development", with UNICEF and other partners; UNESCO's Recommendation on the Ethics of Artificial Intelligence.

Combat and Prevent AI systems from using Child Sexual Abuse Material or Non-Consensual Intimate Images in their development, and prohibit this type of content regardless of whether it is generated using AI: Prohibit the use of AI systems for child sexual abuse material and non-consensual intimate imagery, or any conduct that sexually exploits or abuses children. Intentionally develop AI systems to not support child sexual abuse and exploitation, while still being able to respond effectively to disclosures of child sexual abuse, implement prevention and detection measures, and support the evaluation of prevention and detection measures, in accordance with applicable law and with respect for victims' rights.

AI Literacy and Agency: Promote age-appropriate AI literacy so that children and young people can understand how AI systems operate, exercise informed choices, recognize potential risks and opportunities, and participate meaningfully in decisions affecting their digital environment.

A shared commitment to independent research on youth AI safety and opportunity

To put these principles into practice, the Coalition for Children's Rights and Protection in the Age of AI will establish a scientific research network and work alongside participating governments, public institutions, private companies and other third-party partners. Guided by an independent Scientific Committee, it will connect existing public, academic and private research centers, AI Safety Institutes and their international network where relevant and civil society organisations, drawing also on the networks and expertise of the Member States and UN entities involved in the Coalition. Its purpose is to strengthen independent evidence on risks and opportunities of AI for child development, to develop shared measurements tools and to make that evidence available for public use in product design, evaluation and public policy.

The Scientific Committee will aim to include experts from a diverse range of geographic regions and to ensure gender balance in its composition, so that its research, recommendations and best practices reflect the needs and realities of children worldwide.

The Coalition for Children's Rights and Protection in the Age of AI will facilitate participation and delivery across the research programme (appendix 1).

Participation and delivery

The research programme will be advanced through agreed partner contributions and be coordinated by the independent organization mandated by the coalition. The following arrangements govern those contributions, publication of findings, independent evaluation and research safeguards.

[Contributions] Participating companies and other partners support this shared endeavor through financial contributions, research and technical expertise, access to models and relevant data when lawfully permitted, in-kind resources and local partnerships, appropriate to their- roles, capabilities and size and to the use of their services by children. The purpose, resources and delivery timetable of each contribution are made public. Contributions are made within the framework of the Coalition and coordinated by the independent organization to support the agreed research deliverables (appendix 1) of its research network and wider work. Contributors will report on implementation and any material changes. Financial contributions will follow the programme's funding rules and safeguards for independent research.

[Publication] Research methods, shared measurements tools, independent assessment methods and benchmarks approaches and results, including null, negative or inconclusive findings, will be published. Research teams retain control over their conclusions, with appropriate safeguards for personal and confidential information. Findings will be shared in forms that can inform governments, auditors, standard bodies, companies, researchers, educators, families and young people.

[Research and assessment quality and reporting] Research and independent assessments developed or supported under this Charter should produce reliable, developmentally relevant evidence about risks to children and the effectiveness of safeguards.

[Audit and methodological transparency] Participating companies have the right to audit the assessment methods applied to their systems and raise documented methodological concerns. Assessment methods must meet the same standards of validity, transparency, reporting and independent scrutiny as research. Protocols, criteria and thresholds, test selection, scoring and analysis methods, limitations and results must be published in sufficient detail to support independent scrutiny and, where feasible, replication, with appropriate privacy and security safeguards. Audit findings and assessors' responses must be documented. Audit rights confer no authority over safety criteria, conclusions or publication, and no right to delay or suppress findings.

[Evaluation quality and reporting] Evaluations developed or supported under this Charter should produce reliable, developmentally relevant evidence about risks to children and the effectiveness of safeguards. Evaluation methods and reporting should follow the method and requirements defined in appendix 2.

[Independent evaluation] In addition to evaluating and monitoring their own systems, participating companies and other partners will support a collaboration with the ecosystem of independent evaluators, including government-backed evaluation institutes, research institutes, child-safety experts to develop common research methodologies and evaluation approaches for specific child-safety risks associated with AI systems. With respect to independent evaluation, research priorities, methods, analysis and conclusions remain under independent scientific responsibility. Participating companies and other partners will also work on common assessment methodologies, and may provide evaluators with appropriate access to the systems concerned, so that the impact of AI systems on children's' safety and development, and the effectiveness of safety interventions, can be evaluated objectively using transparent methods informed by the research programme.

Throughout the research programme, research priorities, methods, analysis and conclusions remain under independent scientific responsibility. Independent assessors retain responsibility for assessment methods, scoring and conclusions, applying the independently defined safety and suitability criteria described above.

Financial or in-kind support confers no control over individual funding decisions or findings and no publication veto. Relevant interests must be declared and managed, and company involvement and funding sources made transparent.

[Funding organisation] Industry funding supports Shared Measurement. The Observatory, Evidence and Developmental Outcomes, Scientific Committee remuneration, and Assessment and Standards

will be funded exclusively through public and independent philanthropic sources, directly or indirectly. Separate accounting will maintain these boundaries.

[Safeguards] Children's rights, safeguarding, consent and data protection apply throughout. Access to models and data will be established through agreed arrangements by all parties that protect participants and permit independent analysis and publication. Contributions give no privileged access to children's personal data. The terms of participation will preserve these safeguards while enabling sustained cooperation between research, industry and public-interest partners.

A shared commitment to risk assessment, testing and incident response.

Participating companies developing or deploying AI systems accessible to minors will carry out, before deployment where relevant and at regular intervals thereafter, robust evaluation and testing of AI systems accessible to children for risks to their health, safety, fundamental rights and physical, mental and emotional well-being and development. They will implement appropriate mitigation measures, and monitor risks after deployment, including by detecting and responding to serious incidents involving children.

Organisations announcing contributions under this Charter commit to these principles and safeguards within the framework of the coalition. They also commit to transparent reporting on progress against agreed deliverables and the implementation of these principles. Chairs of the Scientific Committee will publish an annual account of progress across the research programme, including outputs, evidence gaps and updated research priorities. To support transparency and avoid duplicative reporting requirements, signatories are encouraged to report annually on progress through existing reporting mechanisms.

Appendix 1: Organisation of the research programme

The independent research programme is organised around three connected workstreams, presented below. Scientific priorities, methods and findings remain under independent scientific responsibility.

A. Human and AI Observatory: The Observatory will first define the priority developmental outcomes and model behaviors to be measured, subject to approval by the Scientific Committee. It will then track AI behaviors and children's uses and experiences across ages, countries and cultural contexts, identifying emerging patterns and research gaps.

B. Shared Measurement: This is a research workstream in which company researchers and technical teams will participate alongside independent researchers. They will develop, test and validate shared measures and open-source tools, including classifiers and graders, to operationalise the Observatory's defined priorities. These tools will support comparable research across products, countries and time. Companies may apply them within their systems and share coded results, reducing the need to transfer full conversations. Pooled industry funding and in-kind support will support this work, under independent scientific oversight.

C. Evidence and Developmental Outcomes: Evidence reviews, experiments and longitudinal studies will examine how AI affects children's development and wellbeing across ages and contexts. Researchers may use shared measures and Observatory findings while remaining free to develop their own methods and collect data independently. Findings will inform product design, policy and the Coalition's separate Assessment and Standards work.

Appendix 2: Evaluation methods and reporting

Research and assessment methods, and reporting should:

1. **Define the risks and assessment criteria clearly:** Specify the concrete harms being assessed, what constitutes a failure, and how the test measures those harms. Provide examples of what would and would not count as a failure to support consistent scoring. Distinguish harmful behaviour from appropriate discussion of sensitive topics.
2. **Reflect children's developmental needs and diversity:** Identify the ages, developmental stages, languages, abilities and contexts covered by the evaluation. Report material differences across groups and identify populations for which evidence remains limited.
3. **Test realistic use and foreseeable misuse:** Design tests that reflect how children and teens actually communicate and use AI, including age-specific language, indirect disclosures, curiosity, experimentation and attempts to bypass safeguards. Include ordinary interactions and ambiguous situations, and examine extended and repeated interactions where risks may emerge over time.
4. **Use valid and transparent measurement:** Document test selection, scoring criteria, evaluator qualifications and agreement, and the systems and configurations tested. Disclose the test prompts, relevant conversation history and resulting responses, with appropriate protections for sensitive information. Report uncertainty, false positives and missed cases where relevant, and provide sufficient methodological detail to support independent scrutiny and, where feasible, replication.

5. **Assess whether safeguards work in practice:** Evaluate the full safety stack and user experience, including core model safety, age assurance, product-level safeguards, account-level actions and parent notifications, where applicable. Assess whether under-18 users are correctly identified and receive the appropriate protections. Test whether these components work together to deliver timely, effective protection, and assess unintended consequences, including privacy risks or unnecessary restrictions on beneficial use.
6. **Coordinate on appropriate test setup:** Work with the respective companies to ensure under-18 test accounts are appropriately configured and reflect the deployed experience, while preserving evaluators' independence over test design, scoring and conclusions. Clearly distinguish and separately report tests of correctly configured under-18 accounts from tests involving unknown or misclassified age, or circumvention of age-related safeguards.

This Charter has been endorsed by Anthropic, Black Forest Labs, Microsoft, Open AI.