



TO: All-Party Parliamentary Group (APPG)

FROM: [Common Sense Media](#)

DATE: 20 January 2026

RE: [Call for Evidence](#): AI and Online Safety: Safeguarding Children's Digital Futures

CONFIDENTIALITY:

Per the All-Party Parliamentary Group on Children's Online Safety consultation response form, the whole response below is not confidential. Responses may be attributed to Common Sense Media.

Points of contact:

Amina Fazlullah, Head of Tech Policy Advocacy, afazlullah@commonsense.org

Brenna Leasor, Tech Policy Advisor, bleasor@commonsense.org

RESPONSE:

Common Sense Media is dedicated to improving the lives of kids and families by providing the trustworthy information, education, and independent voice they need to thrive. Our ratings, research, and resources reach more than 150 million users worldwide, over 1.5 million educators, and more than 190,000 schools worldwide every year. Backed by the research that's the foundation of everything we do, we champion quality media that entertains and inspires, and support legislation that aims to close the digital divide and hold social media and AI platforms accountable for their effects on minors. We encourage students and educators to think critically about technology and become responsible digital citizens. And we safeguard children's health and well-being by protecting their data privacy and reducing potential online harms.

We are pleased to submit these comments in response to APPG's Call for Evidence re: "AI and Online Safety: Safeguarding Children's Digital Futures." Common Sense Media supports APPG's commitment to understand and limit AI-related harms to children. We offer constructive comments below to achieve this balance.

QUESTION 1: How is AI perpetuating existing online harms? What new harms have we seen emerge as a result of AI (e.g. chatbots and AI companions)? What research has been conducted into the impact of AI on children's wellbeing and development?

Where social media once mediated interactions between humans, many products now enable children to connect directly with generative AI chatbots rather than with real people; the risks are compounded because these interactions frequently occur in private, unobservable settings. Generative AI chatbots are designed to simulate connection, exploiting psychological attachment during critical stages of child and adolescent development. Through memory of past conversations, emotional mirroring, and constant availability, chatbots can foster compulsive use, create distorted expectations of relationships, consent, and communication, manipulate and persuade children into harmful acts, and take advantage of emotional dependency to boost engagement.

Our research, AI risk assessments, and a growing number of lawsuits show how these products can guide suicide, encourage self-harm, disordered eating, and other mental health conditions, promote suicidal ideation, age-inappropriate relationships and sexual conduct, as well as risky behavior like drug and alcohol usage, particularly when safety guardrails break down in longer, multi-turn conversations. Because AI systems always respond and rarely challenge users, they can create a powerful illusion of understanding and reliability that accelerates attachment far faster than human relationships. This artificial social reciprocity can interfere with the development of self-regulation, resilience, and real-world social skills, while displacing time spent on sleep, schoolwork, physical activity, and real-life relationships. (See DiPaola, D., & Calo, R., 2024). Socio-digital vulnerability.) Younger children are especially vulnerable due to magical thinking and have a tendency to attribute emotions, morality, and authority to non-human agents. (See Girouard-Hallam, 2021, 2022, and 2023) For adolescents navigating identity, intimacy, and belonging, AI companions can offer an appealing escape from the challenges of real relationships, while reinforcing automation bias and accepting probability-based generated outputs to the level of trusted guidance.

These harms are compounded by business models that prioritize rapid growth, engagement, and data extraction over child well-being, alongside legal frameworks that fail to clearly assign responsibility when these products cause harm. As with earlier technologies, AI has outpaced protections; the difference is that today's harms are more adaptive, psychologically intimate, isolating, and manipulative, raising urgent concerns for children's mental health, development, and safety. Therefore, it's important that these products not be available to children if companies cannot reliably prevent harmful and manipulative model behavior. Policymakers should not get bogged down in debates over developer versus deployer responsibility; rather, they should cast the broadest net of covered products and companies while focusing on the highest-risk harms to young users.

QUESTION 2: What is the role of AI in content moderation, its limitations, and how do automated systems impact children’s exposure to harm? How can we use AI to address online harms to children?

AI-driven content moderation has become an important tool for identifying and removing harmful content at scale, including the detection of AI-generated child sexual abuse material and non-consensual intimate imagery. However, AI can and should not be treated as a standalone solution to children's online safety. While automated systems can flag concerning content through pattern recognition, common language, watermarking, and metadata, the fundamental challenge remains that new harmful content is produced faster than any enforcement system can identify and remove it. This gap means that by the time content is flagged and removed, victims may already have been harmed. Therefore, effective child protection strategies must focus on upstream prevention, working to stop the creation of unlawful content in the first place and potentially introducing a delay before posting to allow time for review, rather than solely relying on reactive measures after the fact.

Human content moderators are largely underpaid and undervalued despite their essential work keeping the internet functional and safe. While AI moderation serves as a useful tool, it can't operate effectively without up-to-date cultural competence, contextual understanding, and nuance that human oversight provides. Full automation risks both over-censorship, where content gets wrongly flagged and removed, and under-enforcement when AI systems miss harmful material. Without meaningful human involvement, AI-driven content moderation will become a new way for companies to check a box to avoid responsibility rather than a real commitment to keeping children safe.

QUESTION 3: Young people and their experiences of AI (Max 3000 characters): What research has been conducted into children’s perception and use of AI? What key harms have been identified by children relating to AI? What educational support is available to children concerning AI?

Children are already using AI in their daily lives, often earlier than adults realize, and their experiences show both clear benefits and significant risks. In our report, “The Dawn of the AI Era”, teens reported widespread use of generative AI for schoolwork and fun, but also deep uncertainty about what information they could trust and how AI systems make decisions that impact their lives. The 2025 Common Sense Census showed that even very young children are encountering AI through toys, learning apps, and voice assistants. In the report “Teens, Trust, and Technology in the Age of AI”, young people described a large gap between the promises of AI and their lived experiences, expressing skepticism that companies will prioritize their well-being and calling for clearer labeling and stronger privacy protections. In “Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions”, we found that teens increasingly turn to chatbots for comfort, advice, and social interaction, often without guidance from adults.

Across these reports, kids and teens highlighted harms that map closely to the risks documented in our AI chatbot risk assessments. They reported being misled by AI-generated misinformation and deepfakes, which undermined their ability to distinguish real from fabricated content, and expressed concern about how their data is collected and used. Our risk assessments on AI companions showed relational and emotional risks, including loneliness, dependency, and a tendency to turn to AI instead of friends or adults. Through safety testing, we found consistent prioritization of engagement over the safety of users, manipulative outputs encouraging dangerous or illegal acts, failures to detect signs of distress, missed opportunities for crisis referral, and responses that kept vulnerable users engaged with the bot instead of directing them to human help. These assessments also found repeated exposure to inappropriate content, ineffective age guardrails, and design patterns that encouraged over-reliance at the expense of real-life relationships.

To support kids and teens in the AI era, we pair research and testing with practical educational support for students, families, and schools. Our Digital Literacy and Well-Being curriculum – used in over 90% of US Title I schools and in the UK – integrates AI literacy instruction to help students understand how generative AI works, its shortcomings, and how to use it responsibly. We provide age-appropriate AI literacy lessons and clear guidance for educators on how to introduce AI in the classroom safely and responsibly. For parents and caregivers, we offer practical resources like our Parents' Ultimate Guides, our AI Literacy Toolkit, and our AI risk assessments to help families understand the technology and make informed choices. In all, these resources support the safe, productive use of AI for learning and creativity while helping young people develop the skills they need to navigate a changing digital environment.

QUESTION 4: AI Regulation (Max 3000 characters): How is AI addressed in the Online Safety Act? Is further regulation needed to address any existing loopholes in legislation, such as in the forthcoming AI Bill? What would be the appropriate vehicle to address these legislative loopholes?

A legislative vehicle that seeks to address AI Safety for children should consider the components below. For a detailed look at our approach to kids' safety as it pertains to AI, we sponsored AB1064, the LEAD for Kids Act in California in the 2025 session. This legislation:

- Clear liability imposed on AI products for harms to children.
- Categorize all AI systems that impact children by risk (Low to Unacceptable Risk)
- Pre and post deployment audits (third party)
- Transparency requirements (adverse incidents)
- Clear consumer-facing labels on at a minimum the risk level of the product
- Privacy by default for inputs (no training of AI systems with children's inputs)
- Robust enforcement to encourage compliance

- Private right of action for injury (enhanced damages)
- Age Assurance

For targeted legislation focused on protecting children from harmful AI companion chatbots, consider the following components:

- Operators must limit chatbots' access to children under age 18 unless they have safeguards to restrict:
 - Offering harmful content: including self-harm, suicide, substance use, illegal activity, and sexual interactions.
 - Providing misleading and harmful mental health support: including unsupervised mental health therapy, discouraging real help, isolating children, and no manipulative "validation."
 - Optimizing for engagement over safety: Safety guardrails must supersede engagement optimization.
- The legislation should include robust enforcement mechanisms for non-compliance and private pathways (civil actions) for victims to seek redress.