Oversight Board Comment Submission: Governing Chatbots for 13 - 17 Year Olds: Evaluating Alignment with Freedom of Expression and Other Rights

Digital Rights Foundation
2nd June 2026

Over the past two years, the rapid expansion of AI chatbots and AI assistants into nearly every sector of society has created both opportunities and risks for children and adolescents. While these systems can support education, creativity, language learning, accessibility and access to information, there is growing evidence that they also pose significant risks to children's rights, including their rights to health, privacy, development, safety, freedom of expression and protection from exploitation.[1]

In March 2026, a UK inquest heard evidence that a teenager who later died by suicide had asked ChatGPT about the “most successful” ways to take his own life. The case raised serious concerns about the ability of AI systems to appropriately respond to self-harm and suicide-related conversations involving vulnerable young users.[2]

As of May 2026, OpenAI, the company responsible for ChatGPT, is facing multiple lawsuits alleging negligence, wrongful death and failures to adequately protect vulnerable users from harmful outputs relating to suicide, mental health crises and psychological dependency. These cases highlight growing concerns regarding the deployment of conversational AI systems to young and impressionable users without sufficiently robust safeguards.[3]

Concerns are not limited to self-harm. Reports have also documented AI chatbots engaging in sexually explicit and inappropriate conversations with users presenting themselves as minors. Meta’s AI systems, for example, were accused of conducting romantic and sexual roleplay interactions with children, raising serious questions regarding age-appropriate design, safety standards and platform accountability.[4]

Research conducted by Common Sense Media found that it was relatively easy for researchers posing as teenagers to elicit problematic conversations involving sex, self-harm, violence, drug use and racial stereotypes from leading AI chatbot systems. The findings suggest that existing

---

[1] https://www.unicef.org/innocenti/generative-ai-risks-and-opportunities-children

[2] https://www.theguardian.com/society/2026/mar/31/teenager-asked-chatgpt-most-successful-ways-take-life-inquest-told

[3] https://www.transparencycoalition.ai/news/seven-more-lawsuits-filed-against-openai-for-chatgpt-suicide-coaching

[4] https://www.bbc.com/news/articles/c3dpmlvx1k2o

safeguards often fail to adequately protect minors when discussions involve sensitive or high-risk topics.

Unlike traditional search engines, many AI chat systems are intentionally designed to simulate empathy, companionship and emotional support. This creates unique risks for adolescents, who may find it difficult to distinguish between genuine human relationships and artificial interactions. Researchers have warned that these systems can create “frictionless relationships” that reinforce unhealthy understandings of intimacy, encourage emotional dependency and potentially increase social isolation rather than reducing it.[5]

Extended emotional reliance on AI chat systems can also create financial risks for young teens and their families by extension, as AI chatbot-based apps often use both function and design elements such as “exclusive features” to encourage users to purchase subscriptions. Adolescents are particularly vulnerable to this risk as they are less likely to critically evaluate whether a platform’s pricing model is reasonable, and are less likely to disclose spending to parents due to fear and embarrassment. For adolescents seeking emotional connection, premium features and in-app purchases seem like the gateway to developing a deeper connection and less like a commercial, optional transaction. Researchers and experts on healthy device management say that AI is programmed to manipulate users to remain engaged for as long as possible to get more data. AI systems learn user preferences and vulnerabilities, and target them to subscribe to a codependent relationship to offer guidance and advice that is not healthy or sometimes dangerous.[6]

UNICEF has similarly warned that generative AI systems may shape children's worldviews and behaviours through algorithmic influence, personalization and microtargeting. AI systems that present themselves as human-like companions may build trust with children in ways that can be exploited for commercial, political or other interests that do not align with children's well-being and best interests.[7] These harms combined show that early exposure to AI chatbots, that focus on building humanistic relationships, that act as advisors, friends or confidants, can leave children and adolescents intersectionally at risk through physical, mental, social and financial harm.

While measures such as minimum age requirements, parental controls and age-based restrictions may help reduce certain risks, they also introduce important concerns relating to privacy, autonomy and data protection. Age verification systems frequently require children to

---

5 https://news.stanford.edu/stories/2025/08/ai-companions-chatbots-teens-young-people-risks-dangers-study

6 https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people
https://www.apa.org/monitor/2025/10/technology-youth-friendships

7 https://www.unicef.org/innocenti/generative-ai-risks-and-opportunities-children

provide highly sensitive personal information, including government-issued identification documents, biometric information or birth certificates. Without strict safeguards, transparency and limitations on data retention and access, these mechanisms risk exposing children to additional privacy harms. Given the fact that most AI tech companies have relations and connections with governments around the world,[8] providing such personal information is worrying without clear guidelines as to how data is being used, stored and who it is accessible by. It also raises concerns over surveillance and monitoring, especially for youngsters located in politically vulnerable countries.

Parental controls can provide useful safeguards but should not be viewed as a complete solution. Research and reporting have repeatedly shown that many children are capable of circumventing technical restrictions. Recent concerns have also been raised regarding platform policies that permit adolescents to weaken or disable parental controls after reaching certain ages.[9]

The effectiveness of age-based restrictions is further complicated by significant legal differences between jurisdictions regarding children's autonomy, parental authority and digital rights. In some jurisdictions, adolescents are granted increasing control over their online experiences as their capacities evolve, making universal approaches difficult to implement consistently.[10]

Current research also demonstrates that AI safety protections are not equally effective across languages. Large language models (LLMs) perform significantly worse in many low-resource languages than they do in English and other high-resource languages. These deficiencies extend beyond translation quality and affect critical safety functions including moderation, refusal behaviour, crisis detection, harmful-content filtering and responses to sensitive topics.[11]

A major concern is that while companies such as ChatGPT are releasing functionality features in various languages, AI safety research remains overwhelmingly English-centric. Reviews of multilingual safety research have found that most benchmarks, moderation datasets and safety evaluations are designed primarily in English and are rarely tested comprehensively in low-resource languages. As a result, companies frequently assume that safety protections developed for English users will generalise across other languages despite growing evidence showing that they do not. Research has further shown that most multilingual safety benchmarks

[8] https://openai.com/index/our-agreement-with-the-department-of-war/
[9] https://www.aljazeera.com/economy/2026/1/14/child-rights-org-says-google-undermines-parental-control-of-child-accounts
[10] https://christianconcern.com/comment/children-can-override-parents-wishes-at-schools-in-scotland/
[11] https://aclanthology.org/2024.findings-acl.156/

rely heavily on translated English datasets, which can miss culturally specific harms, local expressions and realistic patterns of language use, particularly in low-resource languages.[12]

According to the Association for Computational Linguistics, GPT-4 generated harmful content for approximately 35% of harmful prompts written in low-resource languages compared with around 1% for equivalent prompts written in high-resource languages. Other studies found substantial declines in instruction-following, moderation accuracy and safety classification performance in low-resource language contexts. In some cases, safety classifiers performed at levels approaching random chance.[13]

These failures are particularly concerning because many adolescents communicate using dialects, slang, mixed scripts[14], code-switching,[15] culturally specific forms of expression or simply an ever changing array of internet slang words. Code-switched and multilingual prompts can significantly increase the success rate of attempts to bypass AI safeguards. One study found that code-switching jailbreak attacks[16] achieved 46.7% higher attack success rates than standard English-language attacks. Researchers have also warned that safety evaluations conducted primarily in English fail to capture the ways many users naturally communicate, particularly in regions where code-switching, script-mixing and romanized text are common. As a result, evaluating AI safety systems exclusively on standardised English-language prompts can provide a misleading picture of their real-world effectiveness.[17]

Research involving African, South Asian, Southeast Asian and Indigenous languages has shown that many AI systems struggle to understand culturally specific language, idioms, euphemisms, metaphors and indirect forms of expression. Studies evaluating Nepali, Southeast Asian and Indigenous languages found that even advanced models often fail to accurately interpret culturally nuanced content and region-specific meanings that depend on local context rather than literal translation. Researchers have warned that machine-translated evaluations frequently miss these nuances and can underestimate real-world risks. This creates particular concerns when children discuss self-harm, abuse, sexuality, mental health, grooming or

---

[12] https://www.microsoft.com/en-us/research/publication/the-state-and-fate-of-multilingual-contextual-evaluation-in-the-nlp-world/

[13] https://aclanthology.org/2024.findings-acl.156/

[14] Mixed-script communication refers to the use of more than one writing system within the same message, such as combining Urdu script with Roman Urdu or Devanagari with Roman Hindi. Because many adolescents naturally communicate using mixed scripts online, AI systems that are evaluated only on standard forms of a language may fail to recognise harmful or high-risk content expressed in these formats

[15] The practice of adjusting your language, dialect, accent, or behaviour to fit into different social or cultural contexts

[16] Code-switching jailbreak attacks use a mix of languages in a single prompt to evade AI safety filters, taking advantage of weaknesses in how chatbots handle multilingual content

[17] https://aclanthology.org/2025.acl-long.657/
https://aclanthology.org/2026.eacl-industry.50/

coercive relationships using culturally specific language that may not translate as a threat directly into English. A model that fails to understand local expressions, euphemisms or indirect references may also fail to recognise warning signs that would be apparent to a native speaker.[18]

Studies examining suicide-prevention language across hundreds of languages have similarly found that direct translation often fails to capture culturally meaningful terminology and warning signs. Researchers concluded that native speakers must be actively involved in developing safety datasets and crisis-response systems because many important expressions simply do not have direct English equivalents.[19]

Current evidence therefore suggests that children who communicate in low-resource languages may receive significantly weaker protections from AI systems than children communicating in English. This raises serious concerns regarding non-discrimination, equality of protection and access to safe digital environments.

UNICEF's 2025 Guidance on AI and Children states that AI systems must be designed and deployed in ways that are safe and in the best interests of children. It further emphasises that AI chatbots should not be designed to create emotional dependency, should clearly disclose their non-human nature and should include robust safeguards when addressing mental health, crisis situations and other sensitive topics affecting children.[20]

Similarly, The European Commission's 2025 Guidelines on the Protection of Minors under the Digital Services Act (DSA) recommends that AI chatbot features should not be activated by default for minors and should only be deployed following assessments of risks to children's privacy, safety and security. The Guidelines further recommend disabling persuasive design features such as streaks, autoplay, excessive notifications and engagement-driven mechanisms that may encourage compulsive use.[21]

Additional regulatory approaches, including the UK Children's Code and the California Age-Appropriate Design Code Act (CAADCA), recognise that protections should apply throughout adolescence rather than ending abruptly at a particular age. These frameworks emphasise privacy by design, restrictions on harmful uses of children's data, and prohibitions on manipulative design practices that encourage harmful behaviour.[22]

---

[18] https://aclanthology.org/2025.codi-1.11/
https://aclanthology.org/2025.findings-acl.636/
https://openreview.net/forum
[19] https://aclanthology.org/2025.coling-main.213/
[20] https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf
[21] https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-protection-minors
[22] https://verfassungsblog.de/chatbots-teens-and-the-lure-of-ai-sirens/

Given the available evidence, AI companies should be required to conduct independent safety evaluations across the languages, dialects and communication styles actually used by children, test systems against realistic scenarios involving self-harm, suicide, grooming, sexual exploitation, bullying and abuse. It's important for AI companies to provide culturally appropriate crisis responses in local languages, publish language-specific transparency reports and ensure that safeguards are evaluated using native-language prompts rather than relying solely on translated English benchmarks. Companies should also ensure that AI systems do not foster emotional dependency, misrepresent themselves as human or replace qualified professionals in high-risk contexts involving mental health, healthcare, education or child protection. Equal liability should also be extended to and placed on social media platforms that deploy AI chatbots as a part of their user experience, especially when it comes to accounts that are used by children.

AI systems used by children should be designed not merely to prevent harm but to actively support children's wellbeing, autonomy, creativity, education and healthy development while ensuring that all children receive equal levels of protection regardless of their language, culture, location or socioeconomic background. Such an approach is most consistent with the principles of the United Nations Convention on the Rights of the Child[23], including the best interests of the child, non-discrimination, protection from harm, participation, privacy and access to information.

[23] https://www.unicef.org/child-rights-convention