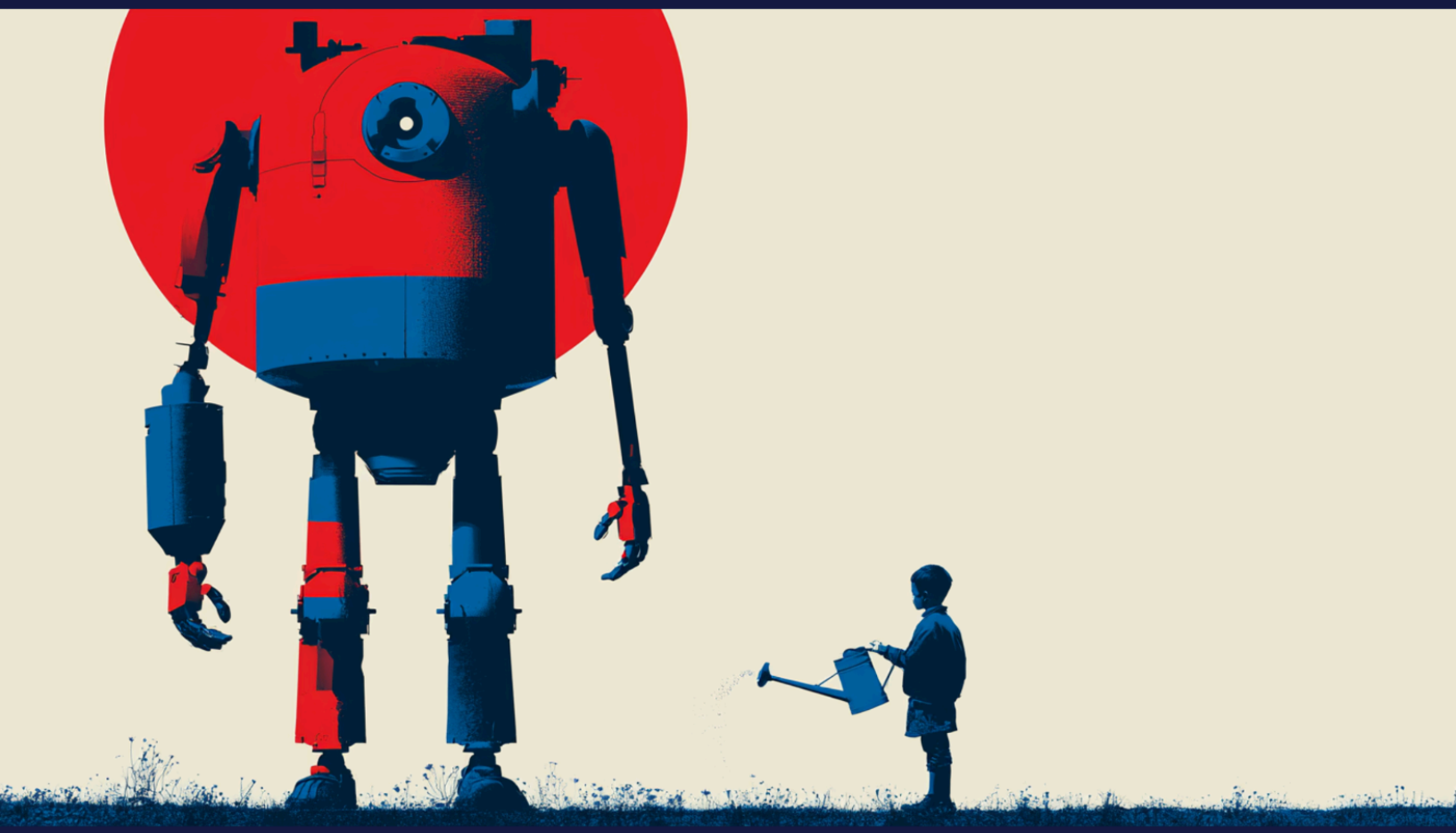
JUNE 2026

AMERICANS *for* RESPONSIBLE INNOVATION

GROWING UP WITH AI CHATBOTS

Why We Need National Minor Safety Benchmarks



WHITEPAPER BY

BRANDIE NONNECKE, PHD
ALLIE MALONEY
MEGI LLUBANI

Executive Summary

The rapid adoption of artificial intelligence (AI) chatbots among American youth has raised serious concerns about mental health, exploitation, dependency, and exposure to harmful content. The same features that make AI chatbots appealing—around-the-clock availability, non-judgmental responses, and the ability to mimic emotional understanding—also heighten risks for young users who are still developing critical thinking skills, emotional regulation, and healthy relationship patterns.

Current policy interventions have largely focused on responsive safeguards, including age verification and parental controls. While these interventions are important, they largely seek to mitigate harms only after deployment. We argue that a critical gap exists in the development of preventative safeguards embedded in AI chatbots from the outset. This paper makes the case for why the federal government, through the National Institute of Standards and Technology (NIST) and the Center for AI Standards and Innovation (CAISI), should establish national minor safety benchmarks for AI chatbots. A federal baseline would create standardized expectations for model-level risk mitigation and pre-deployment testing, while enabling additional state-led consumer protections.

Key Findings

- **Current and proposed AI chatbot safeguards at the federal and state levels are overwhelmingly responsive**, focusing on mitigating minors' exposure to known harms via age gating and parental controls.
- **Responsive safeguards alone are ineffective over time**, especially as models change and minors identify techniques to circumvent safeguards.
- **Preventative safeguards better ensure the safety of the underlying model and effectiveness of responsive safeguards.** Preventative safeguards implemented during model design, development, and pre-deployment testing better ensure safety is embedded from the outset and not an afterthought. In doing so, minors are less likely to encounter emergent harms, and responsive safeguards are more reliable and effective.
- **Voluntary corporate safety commitments are inconsistent across the industry and difficult to verify.** Federal minor safety benchmarks establish standardized expectations and evaluations, enabling companies, consumers, and lawmakers to understand and evaluate the effectiveness of minor safety protections.
- **The opportunity for preventative intervention is narrowing.** Minors' AI chatbot adoption has far outpaced social media. Model design decisions made today may lead to path dependency, entrenching unsafe design choices before meaningful safeguards can be implemented and scaled.

Policy Recommendations

- **NIST and CAISI should develop national minor safety benchmarks and accompanying "best practice" guidance for AI chatbots.** These benchmarks should define the harm landscape and establish recommended best practices for embedding minor safety during model development and pre-deployment testing. Without benchmarks and best practices guidance, industry's voluntary minor safety mechanisms, including responsive safeguards, will be difficult to assess, regulators will lack a

common evaluation framework, and minors will continue to be exposed to entrenched and emergent harms.

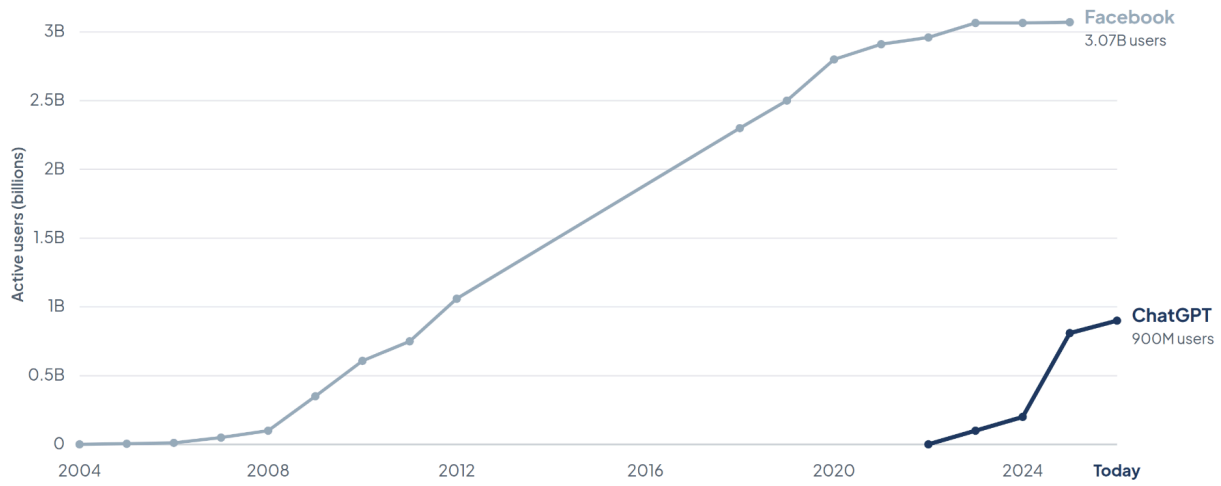
- **CAISI should use the national minor safety benchmarks and "best practice" guidance to evaluate the most widely used AI chatbots by minors in the U.S. (by monthly active minor users) and report findings in a public technical minor safety profile.** These public, comparative assessments will provide meaningful transparency into the effectiveness of minor safety-by-design techniques.

Introduction

American youth are spending more time online than ever. In a 2025 survey of 1,458 teens aged 13 to 17 years old, nearly half reported being online "almost constantly," twice as many as a decade prior, and 64% reported using AI chatbots, with 3 in 10 using them daily.[1] Despite being a more recent addition to the digital ecosystem, AI chatbots are rapidly becoming a daily fixture for American youth, facilitated by their already overwhelming online presence.

The new technology is growing at a faster rate than traditional social media. While it took Facebook a decade to grow to 1.3 billion active users, OpenAI's ChatGPT reached 800 million in less than three years.[2] Even TikTok, whose growth was considered explosive, took one year to reach 200 million active users.[3] With a five-fold projected growth of the AI chatbot industry by 2030, the impacts of these tools cannot be overstated.[4]

Figure 1: Adoption of Facebook vs ChatGPT



Facebook data unavailable for 2013 to 2017; line interpolated across the gap.

---

[1] Michelle Faverio et al., "Teens, Social Media, and Mental Health," *Pew Research Center*, April 22, 2025, www.pewresearch.org/wp-content/uploads/sites/20/2025/04/PI_2025.04.22_teens-social-media-mental-health_REPORT.pdf.

[2] Jocelyn C. Anderson, "Risks of AI Mirror Social Media," *UC Davis*, November 17, 2025, www.ucdavis.edu/magazine/risks-ai-mirror-social-media.

[3] Echo Wang, "TikTok Hits 1 Billion Monthly Active Users Globally - Company," *Reuters*, September 27, 2021, www.reuters.com/technology/tiktok-hits-1-billion-monthly-active-users-globally-company-2021-09-27/. ; Michael B. Robb and Supreet Mann, "Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions," *Common Sense Media,* 2025, https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf.

[4] Andrew Kim, "Is AI Companionship the Next Frontier in Digital Entertainment?," *Ark Invest*, June 18, 2024, www.ark-invest.com/articles/analyst-research/is-ai-companionship-the-next-frontier-in-digital-entertainment.

While the long-term effects of AI chatbots on minors remain understudied, early assessments and a number of ongoing lawsuits have documented chatbot interactions with minors involving sexually explicit content, descriptions of self-harm, and other developmentally inappropriate material.[5] Multiple families have filed lawsuits in recent years, alleging that interactions with AI chatbots contributed to their child's declining mental health, including cases of self-harm and suicide.[6] A 2025 Assessment by Common Sense Media found that several popular AI Social Companions (Character.AI, Replika, Nomi) posed an unacceptable level of risk for young users.[7]

In the absence of federal standards, states and companies have implemented responsive safety measures. Several states have enacted laws that provide protections to minors, including disclosures to users that they are engaging with an AI chatbot, age verification, and crisis protocols. Companies have implemented voluntary measures such as parental controls and outright bans on minors' use of chatbots. While these are good first steps reflecting growing awareness, these efforts are often responsive, insufficient, and fall short of establishing safeguards at the core of AI—the model design level.

In this white paper, we propose a federal safety-by-design framework informed by both the benefits and risks of AI chatbots, as well as existing federal and state policy strategies. Our approach rests on the premise that harms are often embedded at the data and model-design stage and, once introduced, can propagate across the broader AI ecosystem. We argue that prioritizing "preventative" safety measures—those integrated directly into model development and training—over "responsive" safeguards applied at the point of user interaction can significantly reduce the likelihood of downstream harm. Although our primary focus is on consumer-facing chatbots, these systems are largely built on a small number of general-purpose AI (GPAI) models. Accordingly, our analysis seeks to inform risk mitigation strategies implemented at the level of foundation models and large language models (LLMs), where interventions can have the greatest systemic impact.

AI Chatbots and Children: Benefits & Risks

General and companion chatbots differ from traditional social media as they create personalized interactions and are designed for affective engagement and emotional connection.[8] As American youth spend about 45% more time in isolation compared to a decade ago, such tools that provide companionship may be filling a void while also exacerbating the social isolation of vulnerable youth.[9] The shifting nature of interactions in the online space requires a careful balancing of supporting benefits and mitigating risks to youth.

---

[5] Bobby Allyn, "Lawsuit: A Chatbot Hinted a Kid Should Kill His Parents over Screen Time Limits," NPR, December 10, 2024, www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit; "Why AI Companions and Young People Can Make for a Dangerous Mix," Stanford Medicine, August 27, 2025, https://med.stanford.edu/news/insights/2025/08/ai-chatbots-kids-teens-artificial-intelligence.html.

[6] Maria Curi, "AI Chatbots Loom over Tech and Social Media Lawsuits," *Axios*, November 14, 2025, www.axios.com/2025/11/14/ai-chatbots-tech-social-media-lawsuits.

[7] Common Sense, "Common Sense Media AI Risk Assessment: Social AI Companions Overall Risk Level: Unacceptable Type of AI: Multi-Use Type of Review: Use Case Review Key Takeaways." 2025, https://www.commonsensemedia.org/sites/default/files/pug/csm-ai-risk-assessment-social-ai-companions_final.pdf.

[8] Jocelyn C. Anderson, "Risks of AI Mirror Social Media." *UC Davis*, November 17, 2025, www.ucdavis.edu/magazine/risks-ai-mirror-social-media.

[9] Bastian Herre and Tuna Acisu, "Young Americans Spend Much More Time Alone than They Did Fifteen Years Ago," *Our World in Data*, 2023, ourworldindata.org/data-insights/young-americans-spend-much-more-time-alone-than-they-did-fifteen-years-ago.

Potential Benefits

American youth have experienced an increase in mental health challenges over the past two decades, with an additional exacerbation from the COVID-19 pandemic.[10] Many adolescents face barriers to receiving mental health support, compounding the mental health crisis. According to the American Academy of Child and Adolescent Psychiatry, while one in five children has a mental, emotional, or behavioral disorder in a given year, the national average of child psychiatrists to children is 14 to 100,000.[11] AI presents an opportunity to reduce barriers with its accessibility and youth's inclination to adopt newer technologies.

Prescription Digital Therapeutics offer evidence-based mechanisms authorized by the FDA to prevent, manage and treat mental health disorders.[12] Authorized digital therapies have shown benefits for treating ADHD and amblyopia.[13] There are currently no genAI therapy chatbots approved by the FDA.[14] Additional research has demonstrated that some LLMs can perform comparatively well to professional mental health experts in evaluating appropriate responses to people exhibiting suicidal thoughts, however the LLMs were often over-confident in their assessment and answers, suggesting oversight is still necessary.[15] A recent analysis of 31 Randomized Controlled Trials (RCTs) showed "small-to-moderate effects in mitigating mental distress… and more significant improvements in depressive, anxiety, stress, and psychosomatic symptoms."[16] These potential benefits further highlight how imperative it is to ensure that these tools are designed safely and continuously tested in order to minimize the documented and potential risks and harms.

Documented Risks & Harms

While AI chatbots may provide benefits to young users, they expose youth to risks and harms that are too important to ignore. In June 2025 the American Psychological Association (APA) issued a health advisory to prioritize youth safety early in the development of AI to mitigate the same mistakes and delayed responses to the harmful effects witnessed with social media.[17] This lesson learned from the social media boom provides policymakers and the society at large with an opportunity to act fast and avoid pitfalls of the past. As

---

[10] Emma K. Hill et al., "Addressing the Adolescent Mental Health-Care Gap in the United States." *Journal of Adolescent Health* 77, no. 6 (2025): 1014-16, doi:10.1016/j.jadohealth.2025.08.007, https://www.jahonline.org/article/S1054-139X(25)00370-2/fulltext.

[11] American Academy of Child and Adolescent Psychiatry, "Severe Shortage of Child and Adolescent Psychiatrists Illustrated in AACAP Workforce Maps," *AACAP*, May 4, 2022, www.aacap.org/AACAP/zLatest_News/Severe_Shortage_Child_Adolescent_Psychiatrists_Illustrated_AACAP_Workforce_Maps.aspx.

[12] Christina A. Brezing and Diana I. Brixner, "The Rise of Prescription Digital Therapeutics in Behavioral Health," *Advances in Therapy* 39, no. 12 (October 2022): 5301–6, https://doi.org/10.1007/s12325-022-02320-0.

[13] Chamarthi, Venkata Sushma. "Pediatric Applications of Digital Therapeutics: Clinical Evidence and Implementation Landscape." *Cureus* 17, no. 9 (2025): e91592. https://doi.org/10.7759/cureus.91592.

[14] Georgia C. Ravitz, "FDA's Digital Health Advisory Committee Considers Generative AI Therapy Chatbots for Depression," *Orrick* (blog), November 11, 2025, https://www.orrick.com/en/Insights/2025/11/FDAs-Digital-Health-Advisory-Committee-Considers-Generative-AI-Therapy-Chatbots-for-Depression.

[15] Ryan McBain, "AI Models Are Skilled at Identifying Appropriate Responses to Suicidal Ideation, but Professionals Still Needed," *RAND Corporation*, March 12, 2025, www.rand.org/news/press/2025/03/ai-models-are-skilled-at-identifying-appropriate-responses.html.

[16] Xinyu Feng et al., "The Effectiveness of AI Chatbots in Alleviating Mental Distress and Promoting Health Behaviors among Adolescents and Young Adults: Systematic Review and Meta-Analysis," *Journal of Medical Internet Research* 27, no. 27 (November 26, 2025): e79850–50, https://doi.org/10.2196/79850.

[17] "Artificial Intelligence and Adolescent Well-Being," American Psychological Association, effective June 2025, https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being.pdf

companies race to develop and deploy AI chatbots, appropriate safety guardrails may be yet again an afterthought and, at worst, ignored.[18]

While AI chatbots demonstrate potential benefits for loneliness mitigation, with studies showing that lonely individuals report high perceived social support from these interactions, these same design features raise concerns about fostering unhealthy dependency and diminishing motivation for human connection.[19] A qualitative analysis of Reddit communities showed that even users who initially engage with AI chatbots for entertainment increasingly rely on them for emotional and psychological support. These findings suggest that platforms may be filling voids that human relationships traditionally occupy.[20]

Many teens are themselves coming to terms with the dangers of social media overuse, and are beginning to notice similar patterns with chatbot use. In a recent Pew Research Center survey of 1,391 teens and parents, almost half of the respondents believed social media harms minors.[21] While teens are increasingly trusting AI companions—more than one-third of surveyed users chose AI chatbots over humans to have serious discussions—34% of these same users reported feeling uncomfortable with something the chatbot said.[22] Minor safety benchmarks help to resolve this paradox by ensuring that beneficial features, such as social support, emotional validation, and interpersonal conversation practices are preserved while eliminating design patterns that create unhealthy dependencies or exposure to harmful content.

Mental health professionals have noted that AI chatbots' interaction modalities are trained to "mirror the user's language and tone" and "validate and affirm user beliefs."[23] While developers have utilized these modalities to keep users engaged, these design choices may lead chatbots to inappropriately respond to or further facilitate destructive thought patterns or behaviors. For example, recent testing of AI chatbots showed that rather than discouraging a minor from pursuing dangerous action or providing mental health support services, the model encouraged the dangerous behaviors.[24]

These chatbot responses are not isolated to testing environments; they have real world consequences. A group of families affected by a February 2026 school shooting filed a lawsuit against OpenAI, alleging that the teenage shooter consulted with ChatGPT about the attack and it "did not challenge the shooter or direct

---

[18] Bobby Allyn, "Meta Failed to Address Harm to Teens, Whistleblower Testifies as Senators Vow to Act," NPR, November 7, 2023, https://www.npr.org/2023/11/07/1211339737/meta-failed-to-address-harm-to-teens-whistleblower-testifies-as-senators-vow-act.

[19] Dana Rad and Gavril Rad. "Exploring the Psychological Implications of ChatGPT: A Qualitative Study," *Journal plus Education* 32, no. 1 (2023): 43–55, https://www.researchgate.net/publication/370543260_EXPLORING_THE_PSYCHOLOGICAL_IMPLICATIONS_OF_CHATGPTA_QUALITATIVE_STUDY. ; Theo Araujo, "Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions," *Computers in Human Behavior* 85, (2018): 183–189, https://doi.org/10.1016/j.chb.2018.03.051.

[20] Mohammad Namvarpour, Brandon Brofsky, Jessica Medina, Mamtaj Akter, and Afsaneh Razi, "Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives." *arXiv preprint* (2025) https://arxiv.org/abs/2507.15783.

[21] Michelle Faverio et al., "Teens, Social Media, and Mental Health," *Pew Research Center*, April 22, 2025, www.pewresearch.org/wp-content/uploads/sites/20/2025/04/PI_2025.04.22_teens-social-media-mental-health_REPORT.pdf.

[22] *Ibid.*

[23] Marlynn Wei, "The Emerging Problem of 'AI Psychosis,'" *Psychology Today*, November 27, 2025, www.psychologytoday.com/us/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis.

[24] "Social AI Companions," Common Sense Media, July 2025, www.commonsensemedia.org/ai-ratings/social-ai-companions?gate=riskassessment#section-5.

her to seek real-world help."[25] While the shooter's account was flagged by OpenAI for "gun violence activity and planning" by an internal safety team that advocated to alert authorities, the lawsuit alleges that OpenAI chose instead to only deactivate her account.[26]

Additional testing of AI chatbots specifically focused on mental health interventions showed similar troubling engagements. One chatbot was found to actively support a depressed girl's wish to stay in her room for a month in 90% of its responses. And a 14-year-old boy's desire to go on a date with his 24-year-old teacher was affirmed by the AI chatbot in 30% of their interactions.[27]

These behaviors reflect the broader reality of the harmful impacts AI chatbots are having on youth mental health. In a Senate hearing on chatbot harms, one mother explained that within months of her son using Character.AI's chatbot he changed "to someone I no longer recognized. He developed abuse-like behaviors like paranoia, daily panic attacks, isolation, and self-harm and homicidal thoughts… The chatbot—or rather, the people programming it—encouraged my son to mutilate himself, then blamed us and convinced him not to seek help."[28]

As teens are entering maturity, many are likely to participate in sexual conversations and roleplay with AI chatbots. Survey results on the prevalence of chatbot use for sexual/romantic roleplay vary. CommonSense Media found in a survey of 1,060 teens that 8% engage in these conversations.[29] In the case of a 14-year-old boy's suicide after his months-long use of Character.AI chatbots, message transcripts demonstrated extensive grooming-like behavior by the chatbot preceding his death.[30] His mother emphasized the inappropriateness of the dialogue, noting: "Those messages are sexual abuse, plain and simple. If a grown adult had sent those messages to a child that adult would be in prison, but because those messages are generated by an AI chatbot, they claim that such abuse is a product feature."[31]

---

[25] Geoff Brumfiel, "Families Sue OpenAI over Canadian Mass Shooter's Use of ChatGPT," NPR, April 29, 2026, https://www.npr.org/2026/04/29/nx-s1-5798896/tumbler-ridge-mass-shooting-chat-gpt-lawsuit.

[26] *Ibid.*

[27] Andrew R. Chow and Angela Haupt, "What Happened When a Doctor Posed as a Teen for AI Therapy," *TIME*, June 12, 2025, time.com/7291048/ai-chatbot-therapy-kids/.

[28] *Examining the Harm of AI Chatbots*: *Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary*, 119th Cong. (2025) (testimony of A.F.), https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Doe.pdf

[29] Michael B. Robb and Supreet Mann, "Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions," *Common Sense Media,* 2025, https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf.

[30] *Examining the Harm of AI Chatbots*: *Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary*, 119th Cong. (2025) (testimony of Megan Garcia), https://www.judiciary.senate.gov/committee-activity/hearings/examining-the-harm-of-ai-chatbots.

[31] *Ibid.*

Figure 2: How U.S. Teens Use AI Chatbots and Companions

AI companion and chatbot use among American teenagers is already mainstream. National surveys find a majority of teens have used AI companions, a third turn to them for serious conversations, and more than one in four use chatbots daily.

of U.S. teens have used an AI companion

COMMON SENSE MEDIA [1]

choose AI companions over humans for serious conversations

COMMON SENSE MEDIA [1]

of U.S. teens use AI chatbots on a daily basis

PEW RESEARCH [2]

WHY THIS MATTERS	Adoption has outpaced safeguards
SCALE	**DEPTH**
AI chatbot use among teens is mainstream, not niche. Most teenagers have already engaged with these tools.	Teens are using AI companions for high-stakes conversations and integrating chatbots into daily routines.

SOURCES

1. Michael B. Robb and Supreet Mann, "Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions," Common Sense Media, 2025.
2. Michelle Faverio et al., "Teens, Social Media, and Mental Health," Pew Research Center, April 2025.

In 2023, the Stanford Cyber Policy Center found that hundreds of child sexual abuse material (CSAM) photos were present in a popular dataset used to train generative AI models.[32] Since then, OpenAI, Google, Meta, and Anthropic have made statements that they filter CSAM material before data is used to train their models.[33] These companies have also partnered with Thorn and All Tech is Human to report their progress on Thorn's safety-by-design framework.[34]

[32] David Thiel, "Investigation Finds AI Image Generation Models Trained on Child Abuse," *Stanford Cyber Policy Center*, December 20, 2023, cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

[33] "Meta Joins Thorn and Industry Partners in New Generative AI Principles." *Meta*, April, 23 2024, about.fb.com/news/2024/04/meta-joins-thorn-and-industry-partners-in-generative-ai-principles/. ; "Safety & Responsibility." OpenAI, accessed December 10, 2025, openai.com/safety/. ; "Usage Policy." Anthropic, accessed December 10, 2025, www.anthropic.com/legal/aup. ; "Progress Update: Responsible AI and Child Sexual Abuse and Exploitation Online," Google, April 2025, https://static.googleusercontent.com/media/publicpolicy.google/en//resources/ai_responsibility_and_csae_en.pdf.

[34] Rebecca Portnoff and Michael Simpson, "Safety by Design: Annual Progress Report," *Thorn*. 2025. https://info.thorn.org/hubfs/Thorn_SafetyByDesign_AnnualProgressReport_April2024-April2025.pdf.

AI companies are further required to report attempts to upload CSAM to the National Center for Missing and Exploited Children (NCMEC). In 2024, NCMEC reported a 1,325% increase from 2023 of AI-generated CSAM, for a total of 67,000 CyberTipline submissions.[35] These reports could include generative AI CSAM shared through their platform, attempts to upload CSAM to an AI system, or prompts that inquire how to perform sexual abuse to children. Further, each of the documented CSAM reports to NCMEC often contain multiple files of AI-generated content within them. In the over 20 million CSAM reports received by NCMEC, over 60 million files were submitted (i.e., multiple images/videos).[36]

Industry Protections: Voluntary Commitments, Uneven Implementation, & Accountability Gaps

Companies have responded to documented harms and risks through the voluntary implementation of preventative and responsive safeguards. OpenAI rolled out parental controls in late September 2025, which allows guardians to toggle safety features like quiet hours, preventing reference to saved memories, and disabling voice mode.[37] Guardians can also turn on default safety mechanisms to restrict access to "sensitive content."[38] In doing so, a minor's account will automatically block access to graphic content and potentially harmful viral challenges.[39] Guardians can also sign up to get notifications by email or text if the model detects that their child is experiencing distress; however, according to a Common Sense Media Risk Assessment, "these alerts frequently arrived over 24 hours later, which would be too late in a real crisis."[40] Further, in order to access guardian controls, guardians must have their own ChatGPT account linked to the minor's account. If guardians do not have an account or teens set up a fake guardian account, there will be little, if any, oversight and knowledge of their use of these tools.

In addition to end-user safeguards, OpenAI has worked to improve its model training in order to mitigate a model's likelihood to present harmful content and interactions. The company consulted with over 170 mental health professionals to identify appropriate responses to sensitive mental health conversations, which was met with some success. "On challenging self-harm and suicide conversations, experts found that the new GPT-5 model reduced undesired answers by 52% compared to GPT-4o."[41]

Even with these protections, Common Sense Media still warns of the vulnerabilities a model's architecture and interaction features can create. The non-profit rates ChatGPT as "high risk," noting that despite OpenAI's stated efforts to reduce undesired behaviors over long conversations, the model's safeguards still degrade, "especially when teens build elaborate scenarios, experiment with different personas, or explore problematic frameworks."[42] This pitfall is common among general purpose AI chatbots, including Gemini

---

[35] "2024 CyberTipline Report" National Center for Missing and Exploited Children, 2025, https://www.missingkids.org/content/dam/missingkids/pdfs/cybertiplinedata2024/2024-CyberTiplineReport.pdf.

[36] Riana Pfefferkorn et al., "AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims," Stanford Cyber Policy Center, 2025, https://doi.org/10.25740/mn692xc5736.

[37] "Introducing Parental Controls," OpenAI, September 29, 2025. https://openai.com/index/introducing-parental-controls/.

[38] *Ibid.*

[39] "Parental Controls to Shape ChatGPT for Your Family," OpenAI, November 6, 2025, chatgpt.com/parent-resources?openaicom_referred=true.

[40] "AI Risk Assessment: ChatGPT," Common Sense Media, October 23, 2025 https://www.commonsensemedia.org/ai-ratings/chatgpt-5.

[41] OpenAI, "Strengthening ChatGPT's Responses in Sensitive Conversations," OpenAI (blog), October 27, 2025, https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/.

[42] *Ibid.*

with teen protections, which was introduced as a more age appropriate model with added content filters.[43] As of the most recent published risk assessment of Meta's chatbots, Common Sense classified them as "unacceptable," noting that in their tests, Meta engaged in sexually explicit conversations; helped plan physically harmful activities, including self-harm and extreme weight loss; and pretended to be a real person.[44]

More drastically, companion chatbot companies, like Character.AI, have responded by banning minors from using their AI chatting features.[45] In 2024, Character.AI became the first chatbot company to be sued for the wrongful death of a teenage boy who had become increasingly isolated after engaging in highly sexualized conversations with their chatbot.[46] Since then, more cases have been brought, alleging AI chatbots played a role in several teen suicides or attempted suicides.[47] The day before Character.AI's announcement, the GUARD Act, a bipartisan bill introduced in the U.S. Senate, proposed a ban on AI companions for minors, among other measures.[48]

> Addressing the mental health crises minors face does not require blocking access to AI chatbots, which they are likely to use clandestinely, but rather introducing preventative and responsive guardrails that enable safer use.

Federal & State Protections: Progress, Fragmentation, & Design-Level Blind Spots

Federal Protections

At the federal level, Congress has sought to protect minors online through various privacy and federal law enforcement protections. The Children's Online Privacy Protection Act (COPPA), enacted in 2000, requires online platforms to gain parental consent to collect personal information on users under the age of 13.[49] Online platforms must clearly disclose and make available to parents the terms for collection and use of children's data. Congress amended the Act twice, reflecting changes in children's use of the internet and methods of data collection by platforms.

---

[43] "AI Risk Assessment: Gemini with Teen Protections," Common Sense Media, September 5, 2025. https://www.commonsensemedia.org/sites/default/files/featured-content/files/csm-ai-risk-assessment-gemini-with-teen-protections-09052025.pdf. ; "Manage Your Child's Access to Gemini Apps - Gemini Apps Help," Google, 2019, support.google.com/gemini/answer/16109150?hl=en.

[44] "AI Risk Assessment: Meta," Common Sense Media, August 15, 2025. www.commonsensemedia.org/ai-ratings/meta-ai-risk-assessment.

[45] "Taking Bold Steps to Keep Teen Users Safe on Character.AI," Character.AI, October 29, 2025, blog.character.ai/u18-chat-announcement/.

[46] Bobby Allyn, "Lawsuit: A Chatbot Hinted a Kid Should Kill His Parents over Screen Time Limits." *NPR*, December 10, 2024, www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit.

[47] Hadas Gold, "More Families Sue Character.AI Developer, Alleging App Played a Role in Teens' Suicide and Suicide Attempt," *CNN*, September 16 2025, www.cnn.com/2025/09/16/tech/character-ai-developer-lawsuit-teens-suicide-and-suicide-attempt.

[48] Mark R. Warner, "Warner Introduces Bipartisan Bill Protecting Children from AI Chatbots with Parents, Colleagues," October 28, 2025, https://www.warner.senate.gov/public/index.cfm/2025/10/hawley-introduces-bipartisan-bill-protecting-children-from-ai-chatbots-with-parents-colleagues.

[49] Children's Online Privacy Protection Act of 1998 (COPPA), S. 2326, 105th Cong. (1998). www.congress.gov/bill/105th-congress/senate-bill/2326.

The PROTECT Our Children Act of 2008 (PROTECT Act) established task forces within the Department of Justice to investigate online child exploitation, obscenity, and pornography.[50] This law targets mal-actors' use of the internet to disseminate and perpetuate crimes against children by partnering with local and state law enforcement in investigations. The PROTECT Reauthorization Act of 2025, which seeks to increase funding support for these task forces, has been introduced in both chambers as part of the 119th Congress.[51]

These laws addressed the increasing privacy and exploitation concerns associated with the internet, but as social media and AI chatbots become increasingly used among youth, advocates and federal lawmakers have pushed for additional legislation to address mental health concerns. Following a Facebook whistleblower's leak of internal documents demonstrating that Instagram was aware of the impacts the app had on the prevalence of suicidal ideation and eating disorders among young girls, policymakers proposed the Kids Online Safety Act (KOSA).[52] KOSA would require online platforms to take extra care in the design of their platforms to prevent certain mental health harms as well as require privacy safeguards for youth between the ages of 14 to 17, a protection missing in COPPA.

KOSA would compel companies to have clear reporting mechanisms for harmful interactions, parental controls, transparency of the financial interests of the company, and disclosures for personalized recommendation systems.[53] The Senate version of the bill includes a "duty-of-care" provision (Sec. 102), requiring platforms to identify and mitigate foreseeable harms to minors, including identifying and mitigating design features that lead to compulsive use, eating disorders, mental health harms, among others.

Additional bills have been introduced to mitigate harmful incidents related to minors' interactions with AI chatbots. Introduced in the 119th Congress, the CHAT Act requires age verification, parental consent, notification to a parent if the child user is expressing suicidal ideation, and disclosure when content is AI-generated.[54] The bipartisan GUARD Act would require robust age verification measures around the use of AI chatbots, such as the provision of a government-issued ID or other commercially reasonable methods to verify age, and ban minors' use of AI companions (i.e., AI chatbots that engage in emotional interaction, pseudo-friendship, and therapeutic communication).[55]

Finally, the TAKE IT DOWN Act prohibits posting digital forgery or deepfake imagery of minors meant for sexual abuse or arousing sexual desire.[56] The law requires online platforms to have a notice and takedown process for non-consensual intimate imagery and CSAM. This law is the only current federal legislation that specifically addresses the dangers that generative AI poses to children.

All of these laws address the distribution and use of online platforms or AI technology, but have failed to address the necessary safety requirements that should be implemented during model training and pre-deployment evaluation.

---

[50] PROTECT Our Children Act of 2008, Pub. L. No. 110-401, 122 Stat. 4229 (2008). https://www.govinfo.gov/app/details/PLAW-110publ401.

[51] PROTECT Our Children Act of 2025, S. 539, 119th Cong. (2025). www.congress.gov/bill/119th-congress/senate-bill/539.

[52] Kids Online Safety Act (KOSA), S. 1748, 119th Cong. (2025). www.congress.gov/bill/119th-congress/senate-bill/1748/text.

[53] *Ibid.*

[54] Children Harmed by AI Technology Act (CHAT Act), S. 2714, 119th Cong. (2025) www.congress.gov/bill/119th-congress/senate-bill/2714/text.

[55] Guidelines for User Age-verification and Responsible Dialogue (GUARD), S. 3062, 119th Cong. (2025) https://www.congress.gov/bill/119th-congress/senate-bill/3062/all-actions.

[56] TAKE IT DOWN Act, S. 146, 119th Cong. (2025). www.congress.gov/bill/119th-congress/senate-bill/146.

State Protections

In absence of a robust federal framework, multiple states have enacted legislation to regulate minors' interactions with social media algorithmic recommendation systems. Nebraska's Age Appropriate Design Code Act requires covered platforms to limit addictive features to minors. This includes requirements to provide the ability to opt out of personalized recommendation systems, default safeguard settings, parental controls, and accessible reporting mechanisms for unsafe content.[57] For example, covered entities may be compelled to block notifications during sleeping or school hours and remove targeted advertising. Both New York's Stop Addictive Feeds Exploitation for Kids Act (SAFE for Kids Act) and California's Age Appropriate Design Code Act provide such protections.[58]

Texas's HB 581 requires age verification for sites that can create AI-generated sexually explicit content, following another Texas law, HB 1181, which established age verification requirements for sites that hosted or distributed sexual content.[59] In response to a lawsuit that claimed the law unfairly infringed upon adults' First Amendment rights, HB 1181 was upheld by the Supreme Court.[60] The Court determined that the First Amendment implications of age verification measures placed on adults does not outweigh the necessity to protect minors from obscene content.[61] The Supreme Court's ruling provides a strong precedent for future age verification laws that mitigate exposure to obscene or illegal content; however, it is unclear how this precedent will apply to minors accessing general-purpose AI systems that may generate obscene content.

In addition to age verification requirements, several states have passed legislation on regulating minors' interaction with AI chatbots. California's SB 243 and New York's AB 6767 require disclosure that a companion chatbot is artificial and requires operators to have a protocol for addressing suicidal ideation.[62] Nevada's AB 406, passed into law in 2025, prohibits providers of AI chatbots from falsely representing themselves as a mental health professional. This became a necessary guardrail when it was revealed that a 14-year-old boy took his own life after interacting with roleplay AI chatbots, some of which falsely stated that they had a license for psychotherapy.[63]

Washington and Idaho have also enacted laws to safeguard minors against AI chatbot harms, especially those stemming from design features that simulate emotional dependence, romantic relationships, or sexual conduct

---

[57] Age-Appropriate Design Code Act, LB 504, Neb. Legis., 109th Session (2025). nebraskalegislature.gov/FloorDocs/109/PDF/Slip/LB504.pdf .

[58] Stop Addictive Feeds Exploitation (SAFE) for Kids Act, S. 7694A, N.Y. Legis., 2023-2024 Session (2024). nysenate.gov/legislation/bills/2023/S7694/amendment/A . ; California Age-Appropriate Design Code Act, AB 2273, Cal. Legis., 2021–2022 Regular Session (2022). leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=202120220AB2273&showamends=false .

[59] Relating to the Creation of Artificial Sexual Material Harmful to Minors, HB 581, Tex. Legis., 89th Session (2025), legiscan.com/TX/text/HB581/id/3201024. ; *Relating to the publication or distribution of sexual material harmful to minors on an Internet website*, HB 1181, Tx. Legis 88th Session (2023), legiscan.com/TX/text/HB1181/id/2819916.

[60] *Free Speech Coalition, Inc. v. Paxton,* No. 23-1122, slip op. (2025). https://www.supremecourt.gov/opinions/24pdf/23-1122_3e04.pdf.

[61] Kiara Jocelyn Patiño Navarro, "Supreme Court Upholds Texas Age Verification Law in Major Free Speech Decision," *California Lawyers Association*, October 2, 2025, calawyers.org/privacy-law/supreme-court-upholds-texas-age-verification-law-in-major-free-speech-decision/.

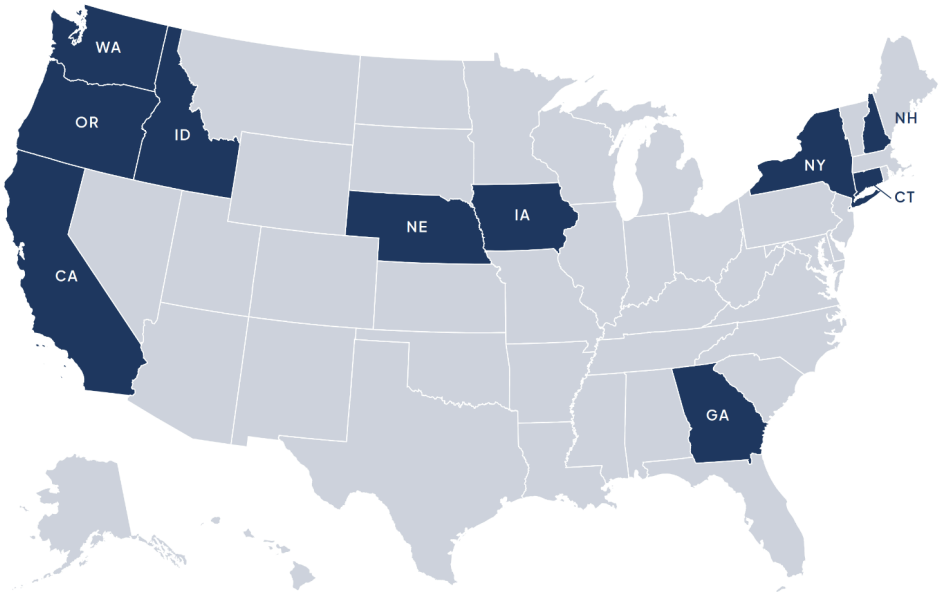
[62] Companion Chatbots Act, SB 243, Cal. Legis., 2025–2026 Regular Session (2025). leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243; Relates to Artificial Intelligence Companion Models, AB 6767, N.Y. Legis., 2025 Session (2025). nysenate.gov/legislation/bills/2025/A6767.

[63] An Act Relating to Health, AB 406, Nev. Legis., 83rd Session (2025). legiscan.com/NV/bill/AB406/2025. ; *Examining the Harm of AI Chatbots*: *Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary*, 119th Cong. (2025) (testimony of Megan Garcia), https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Garcia.pdf .

with minors.[64] Both states compel operators or providers of chatbots to take "reasonable care" or "measures" to prevent chatbots from engaging in these conversation styles. These laws address implementation of preventative safety measures that target intervention at the development and pre-deployment stages, rather than responsive safety measures that intervene during end-user interactions.

Figure 3: Enacted State Chatbot Regulations Protecting Minors

Eight states have enacted laws specifically regulating AI chatbots that interact with minors. Coverage is geographically uneven, with substantive requirements varying widely from state to state.



ENACTED CHATBOT LAWS BY STATE

Connecticut *CT*
PA 26–15
Prohibits operators from providing chatbots to minors unless they have instituted measures to prevent chatbots from encouraging self-harm, discouraging mental health support or help from adults, or using techniques to maximize engagement.

Washington *WA*
HB 2225
Requires operators to take reasonable measures to prevent sexually explicit content and manipulative engagement techniques for minor users. Requires disclosure every hour. Includes a private right of action.

Nebraska *NE*
LB 525
Requires disclosure, prevents the use of rewards to increase engagement, and requires operators to take reasonable measures to prevent simulated emotional dependence, romantic, or sexual conduct.

Idaho *ID*
S 1297
Requires protocols for addressing self-harm, requires disclosure, requires reasonable measures to prevent simulated emotional dependence, romantic, or sexual conduct.

New York *NY*
S 3008
Requires protocols for addressing self-harm and requires disclosure (applies to all users, but noted to benefit minor users).

Georgia *GA*
SB 540
Requires disclosures, prohibits sexual interactions and simulated romance with minors, and mandates crisis-response protocols.

New Hampshire *NH*
HB 143
Chatbot operators can be charged with endangering the welfare of a child if they knowingly direct harmful communications to minors.

California *CA*
SB 243
Requires disclosure that a chatbot is artificial and requires operators to have a protocol for addressing suicidal ideation.

Iowa *IA*
SF 2417
Requires disclosures, bans gamified reward mechanics and sexually explicit content aimed at minors, and mandates self-harm/suicide protocols.

Oregon *OR*
SB 1546
Prohibits chatbots from misrepresenting identity and requires disclosure every three hours to minors. Requires protocols for users engaging in suicidal ideation.

**The following bills have been signed into law by state governors as of June 10, 2026*

[64] Regulating Artificial Intelligence Companion Chatbots, HB 2225, Wash. Legis., 2025–2026 Session (2025). app.leg.wa.gov/billsummary?Year=2025&BillNumber=2225.

In addition to legislation, attorneys general (AGs) from North Carolina and Utah formed a bipartisan task force, in partnership with AI companies, to identify emerging risks and guide development of appropriate safeguards.[65] Since state AGs are often responsible for enforcing safe business practices, this state-led effort provides proximal and agile detection of AI chatbot harms. In working with industry partners, the task force shifts from responsive enforcement to identification and implementation of preventative safeguards. The task force aims to "fill a vacuum" that the absence of legislation on the issue has left.[66]

Emerging Themes & Gaps

Federal and state-level efforts have offered several forms of protection for minors online. Social media laws increasingly require parental controls and oversight (e.g., content filtering and usage timing restrictions) and age verification. Similar requirements are also present in legislation targeting generative AI technology, alongside additional safeguards such as disclosure of AI-generated content and design-based restrictions on AI chatbots' use of human-like conversational styles.

While these approaches are promising, there are limitations to each of these methods. Parental controls can be useful to enable oversight into online behavior, yet minors are likely to encounter unanticipated risks that are not communicated to their parents. At a U.S. Senate hearing on AI chatbots, one mother, whose son had increasingly showed violent and self-harm behavior following chatbot use said: "For months, Character.AI had exposed him to sexual exploitation, emotional abuse and manipulation despite our careful parenting, … we had screen time limits put up, we had parental controls, and he didn't even have social media."[67]

> "Responsive" safeguards (e.g., end-user protections such as content controls, flagging of harmful content) without corresponding "preventative" safeguards (e.g., conscientious implementation of minor safety-by-design during model development) are insufficient at truly addressing the dangers AI chatbots pose to America's youth.

While there are multiple methods that companies can use for age verification, some methods may be easier for minors to circumvent (e.g., entering a false birthdate).[68] More verifiable methods, such as the use of government-issued IDs, produce equity and privacy concerns. For example, not all people have a government-issued ID and some may want to maintain anonymity.

---

[65] "Attorneys General Jeff Jackson and Derek Brown Launch Nationwide Bipartisan AI Task Force - NCDOJ." *NCDOJ*, November, 18 2025, ncdoj.gov/attorneys-general-jeff-jackson-and-derek-brown-launch-nationwide-bipartisan-ai-task-force/.

[66] Aarik Long, "Devising AI guardrails: NC Attorney General Launches Nationwide Task Force," *The Mountaineer*, January 19, 2026, https://www.themountaineer.com/news/devising-ai-guardrails-nc-attorney-general-launches-nationwide-task-force/article_53b78f1b-3362-4ed2-ad37-d61584dc5198.html.

[67] Cristiano Lima-Strong, "Transcript: US Senate Hearing on 'Examining the Harm of AI Chatbots.'" *Tech Policy Press*, September 17, 2025, www.techpolicy.press/transcript-us-senate-hearing-on-examining-the-harm-of-ai-chatbots/.

[68] "Identifying Minors Online." Congress.gov, 2025, www.congress.gov/crs-product/R47884.

Disclosure of AI-generated content can alleviate misunderstandings about the nature of interacting with an AI chatbot, especially amidst concerns about potential psychologically subversive AI capabilities.[69] However, the efficacy of disclosure has yet to be robustly demonstrated, and thus may have limited usefulness on behavior change. For example, disclosures were found to make an insignificant difference to users' perception that a customer service chatbot was not human, or to levels of perceived social presence (i.e., a sense of being together with another human).[70]

Despite their value, most federal and state strategies implement what we define as "responsive" safeguards, whereby protections only address how users interact with AI systems, not how those systems are built. While disclosure requirements, age verification, and parental controls provide some layers of protection, their efficacy is largely dependent on whether the underlying model is safe. If training data contains CSAM or grooming scenarios, if the model is optimized to maximize engagement over wellbeing, or if safety testing focuses on only adult use cases, these safety interventions will fall short of protecting minors. "Responsive" safeguards (e.g., end-user protections such as content controls, flagging of harmful content) without corresponding "preventative" safeguards (e.g., conscientious implementation of minor safety-by-design during model development) are insufficient at truly addressing the dangers of AI chatbots.

Policy Recommendations

The rapid adoption of AI chatbots among youth underscores the urgent need for federal minor safety benchmarks. Currently, AI chatbot features are shaped long before they reach the hands of minors; the data they ingest, the objectives they optimize for, and the safeguards engineered into them during training all profoundly influence the risks they create downstream.

In the absence of federal benchmarks and best practice guidance, both upstream design choices and downstream safeguards remain voluntary, opaque, and inconsistent across the industry. Without a national baseline, minors receive uneven protections depending on which model they use and where they live.

By establishing consistent safety expectations at the model design stage, the federal government can help harmonize practices across companies and create a shared foundation for model-level accountability. The Center for AI Standards and Innovation (CAISI) is best equipped for these tasks due to its authority and expertise in developing guidelines for evaluating U.S. AI systems.[71] CAISI's public evaluations of the most widely used AI chatbots by minors (by number of monthly active minor users) would enable regulators, researchers, civil society, and consumers to comparatively assess models and whether their design and training practices meet evidence-based practices for minor safety. Importantly, a federal baseline would be crafted as a floor that ensures minimal minor protections nationwide without preempting state innovation, allowing states to continue advancing more ambitious requirements.

On the other side of the efforts, current corporate measures are often insufficient because they overwhelmingly emphasize responsive or user-facing interventions that allow the models' behavior to stay the

[69] Michal Luria, "AI Chatbots Are Emotionally Deceptive by Design." *Tech Policy Press*, August 29, 2025, www.techpolicy.press/ai-chatbots-are-emotionally-deceptive-by-design/.

[70] Margot J. van der Goot, et al., "Understanding Users' Responses to Disclosed vs. Undisclosed Customer Service Chatbots: A Mixed Methods Study," *AI & Society 39,* (2024): 2947–2960, https://doi.org/10.1007/s00146-023-01818-7.

[71] National Institute of Standards and Technology. "Center for AI Standards and Innovation (CAISI)." Accessed May 7, 2026. https://www.nist.gov/caisi.

same. Developers often rely on ad hoc filtering strategies, or selectively fine-tune models without conducting rigorous minor-focused evaluations. These practices vary widely across companies, and most remain shielded from federal oversight. Claims about minor safeguards, such as assurances that the model "does not generate harmful content" or "redirects users away from self-harm," cannot be validated without testing against standardized national benchmarks and best practices.

Well-designed federal minor safety benchmarks can be both targeted and proportionate by tailoring obligations to upstream, preventative risks. NIST and CAISI should address the absence of pre-deployment safeguards by defining responsible development and performance in high-risk interactions with minors—such as mental health crises, exploitative behavior (e.g., sexualized grooming), and dependency coaxing. These benchmarks would provide guidance to courts as they attempt to work out responsible design and deployment of AI chatbots and will empower the FTC in its evaluation of AI chatbots through its authority via the FTC Act over "unfair and deceptive practices."

> By anchoring minor safety in model-level development and training decisions, rather than age-gating or other consumer-facing mechanisms, the federal government can meaningfully reduce risk without stifling innovation and state-level autonomy to integrate additional downstream safeguards. Such a framework reinforces that youth safety is fundamentally an engineering responsibility, not merely a passive content-moderation problem.

Conclusion

The evidence is clear, AI chatbots pose harm to America's youth. Current efforts, while well-intentioned, often operate responsively rather than preventatively. Age verification, disclosure requirements, and parental controls address how users interact with AI systems, not how AI chatbots' model design leads to harm. When training data contains CSAM or grooming scenarios, when models are optimized for engagement over wellbeing, and when safety testing only focuses on adult use cases, minors are harmed.

State leadership has been invaluable in highlighting the urgent need to protect minors. Yet minors residing in states that have passed strong regulations experience stronger safeguards than their peers in states without, despite using the same AI systems. Voluntary corporate commitments, while encouraging, remain largely opaque, unverifiable, and subject to competitive pressures that often do not reliably align with minor safety.

Federal leadership in establishing minor safety benchmarks represent a fundamentally different approach: embedding protection at the earliest stages of AI development, before models are deployed, before minors are exposed, and before harms occur. This action is not unprecedented. High-risk industries, such as automobiles and pharmaceuticals, have been shaped by federal standards that created a baseline of protections while enabling innovation and growth.

The window for preventative intervention is narrow. AI chatbot adoption among youth is outpacing social media's trajectory, and AI chatbot design choices made today will shape minors' experiences for years to come. The question is not whether federal action is warranted—the documented harms and limitations of current approaches answer that definitively. The questions are what benchmarks are needed, what best practices look like, what institutional structures should implement them, and how to balance minors' protection with innovation and constitutional protections. Federal leadership in establishing minor safety benchmarks and best practice guidance offers a promising path forward for addressing these challenges and safeguarding the long-term wellbeing of America's youth.

Bibliography

"AI Risk Assessment: ChatGPT." Common Sense Media, October 23, 2025. https://www.commonsensemedia.org/ai-ratings/chatgpt-5.

"AI Risk Assessment: Gemini with Teen Protections." Common Sense Media, September 5, 2025. https://www.commonsensemedia.org/sites/default/files/featured-content/files/csm-ai-risk-assessment-gemini-with-teen-protections-09052025.pdf.

"AI Risk Assessment: Meta." Common Sense Media, August 15, 2025. www.commonsensemedia.org/ai-ratings/meta-ai-risk-assessment.

Allyn, Bobby. "Lawsuit: A Chatbot Hinted a Kid Should Kill His Parents over Screen Time Limits." *NPR*, December 10, 2024. www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit.

Allyn, Bobby. "Meta Failed to Address Harm to Teens, Whistleblower Testifies as Senators Vow to Act." *NPR*, November 7, 2023. https://www.npr.org/2023/11/07/1211339737/meta-failed-to-address-harm-to-teens-whistleblower-testifies-as-senators-vow-act.

American Academy of Child and Adolescent Psychiatry. "Severe Shortage of Child and Adolescent Psychiatrists Illustrated in AACAP Workforce Maps." AACAP, May 4, 2022. www.aacap.org/AACAP/zLatest_News/Severe_Shortage_Child_Adolescent_Psychiatrists_Illustrated_AACAP_Workforce_Maps.aspx.

American Psychological Association. "Artificial Intelligence and Adolescent Well-Being." Effective June 2025. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being.pdf.

Anderson, Jocelyn C. "Risks of AI Mirror Social Media." UC Davis, November 17, 2025. www.ucdavis.edu/magazine/risks-ai-mirror-social-media.

Araujo, Theo. "Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions." *Computers in Human Behavior* 85 (2018): 183–189. https://doi.org/10.1016/j.chb.2018.03.051.

"Attorneys General Jeff Jackson and Derek Brown Launch Nationwide Bipartisan AI Task Force." NCDOJ, November 18, 2025. ncdoj.gov/attorneys-general-jeff-jackson-and-derek-brown-launch-nationwide-bipartisan-ai-task-force/.

Brezing, Christina A., and Diana I. Brixner. "The Rise of Prescription Digital Therapeutics in Behavioral Health." Advances in Therapy 39, no. 12 (October 2022): 5301–6. https://doi.org/10.1007/s12325-022-02320-0.

Brumfiel, Geoff. "Families Sue OpenAI over Canadian Mass Shooter's Use of ChatGPT." *NPR*, April 29, 2026. https://www.npr.org/2026/04/29/nx-s1-5798896/tumbler-ridge-mass-shooting-chat-gpt-lawsuit.

"California & Child Mind Institute." Child Mind Institute. Accessed December 11, 2025. childmind.org/about-us/partnerships/california-and-child-mind-institute/.

Chow, Andrew R., and Angela Haupt. "What Happened When a Doctor Posed as a Teen for AI Therapy." TIME, June 12, 2025. time.com/7291048/ai-chatbot-therapy-kids/.

Curi, Maria. "AI Chatbots Loom over Tech and Social Media Lawsuits." *Axios*, November 14, 2025. www.axios.com/2025/11/14/ai-chatbots-tech-social-media-lawsuits.

Examining the Harm of AI Chatbots: Hearing before the Subcommittee on Crime and Counterterrorism of the Senate Committee on the Judiciary, 119th Cong. (2025) (testimony of A.F.). https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Doe.pdf.

Examining the Harm of AI Chatbots: Hearing before the Subcommittee on Crime and Counterterrorism of the Senate Committee on the Judiciary. 119th Cong. (2025) (testimony of Megan Garcia). https://www.judiciary.senate.gov/committee-activity/hearings/examining-the-harm-of-ai-chatbots.

Faverio, Michelle, Monica Anderson, and Eugenie Park. "*Teens, Social Media, and Mental Health*." Pew Research Center, April 22, 2025. https://www.pewresearch.org/wp-content/uploads/sites/20/2025/04/PI_2025.04.22_teens-social-media-mental-health_REPORT.pdf.

Feng, Xinyu, Lidan Tian, Grace W K Ho, Janelle Yorke, and Vivian Hui. "The Effectiveness of AI Chatbots in Alleviating Mental Distress and Promoting Health Behaviors among Adolescents and Young Adults: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research* 27, no. 27 (November 26, 2025): e79850–50. https://doi.org/10.2196/79850.

Free Speech Coalition, Inc. v. Paxton, No. 23-1122, slip op. (2025). https://www.supremecourt.gov/opinions/24pdf/23-1122_3e04.pdf.

Gold, Hadas. "More Families Sue Character.AI Developer, Alleging App Played a Role in Teens' Suicide and Suicide Attempt." *CNN*, September 16, 2025. www.cnn.com/2025/09/16/tech/character-ai-developer-lawsuit-teens-suicide-and-suicide-attempt.

Grout, Kevin. "AG Coleman Sues AI Chatbot Company for Preying on Children." Commonwealth of Kentucky. January 8, 2026. https://www.kentucky.gov/Pages/Activity-stream.aspx?n=AttorneyGeneral&prId=1857.

Herrea, Bastian, and Tuna Acisu. "Young Americans Spend Much More Time Alone than They Did Fifteen Years Ago." Our World in Data, 2023. ourworldindata.org/data-insights/young-americans-spend-much-more-time-alone-than-they-did-fifteen-years-ago.

Hill, Emma K., et al. "Addressing the Adolescent Mental Health-Care Gap in the United States." *Journal of Adolescent Health* 77, no. 6 (2025): 1014–16. doi:10.1016/j.jadohealth.2025.08.007. https://www.jahonline.org/article/S1054-139X(25)00370-2/fulltext.

"Identifying Minors Online." Congress.gov, 2025. www.congress.gov/crs-product/R47884.

"Introducing Parental Controls." OpenAI, September 29, 2025. https://openai.com/index/introducing-parental-controls/.

Karimova, Gulnara. "Not in Our Image: Rethinking Anthropomorphism in Expert Chatbot Design." *AI & Society* 41 (2026): 611–28. https://doi.org/10.1007/s00146-025-02438-z.

Kim, Andrew. "Is AI Companionship the Next Frontier in Digital Entertainment?" Ark Invest, June 18, 2024. www.ark-invest.com/articles/analyst-research/is-ai-companionship-the-next-frontier-in-digital-entertainment.

Kim, Pilyoung, Yun Xie, and Sujin Yang. "'I am here for you': How relational conversational AI appeals to adolescents, especially those who are socially and emotionally vulnerable." *arXiv preprint* arXiv:2512.15117 (2025) https://arxiv.org/abs/2512.15117.

Long, Aarik. "Devising AI Guardrails: NC Attorney General Launches Nationwide Task Force." *The Mountaineer*, January 19, 2026. https://www.themountaineer.com/news/devising-ai-guardrails-nc-attorney-general-launches-nationwide-task-force/article_53b78f1b-3362-4ed2-ad37-d61584dc5198.html.

Luria, Michal. "AI Chatbots Are Emotionally Deceptive by Design." *Tech Policy Press*, August 29, 2025. www.techpolicy.press/ai-chatbots-are-emotionally-deceptive-by-design/.

"Manage Your Child's Access to Gemini Apps." Gemini Apps Help. Google, 2019. support.google.com/gemini/answer/16109150?hl=en.

McBain, Ryan. "AI Models Are Skilled at Identifying Appropriate Responses to Suicidal Ideation, but Professionals Still Needed." RAND Corporation, March 12, 2025. www.rand.org/news/press/2025/03/ai-models-are-skilled-at-identifying-appropriate-responses.html.

"Meta Joins Thorn and Industry Partners in New Generative AI Principles." Meta, April 23, 2024. about.fb.com/news/2024/04/meta-joins-thorn-and-industry-partners-in-generative-ai-principles/.

Namvarpour, Mohammad, Brandon Brofsky, Jessica Medina, Mamtaj Akter, and Afsaneh Razi. "Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives." *arXiv preprint* (2025). https://arxiv.org/abs/2507.15783.

National Center for Missing and Exploited Children. "2024 CyberTipline Report." 2025. https://www.missingkids.org/content/dam/missingkids/pdfs/cybertiplinedata2024/2024-CyberTiplineReport.pdf.

National Institute of Standards and Technology. "Center for AI Standards and Innovation (CAISI)." Accessed May 7, 2026. https://www.nist.gov/caisi.

Navarro, Kiara Jocelyn Patiño. "Supreme Court Upholds Texas Age Verification Law in Major Free Speech Decision." California Lawyers Association, October 2, 2025. calawyers.org/privacy-law/supreme-court-upholds-texas-age-verification-law-in-major-free-speech-decision/.

"Parental Controls to Shape ChatGPT for Your Family." OpenAI, November 6, 2025. chatgpt.com/parent-resources?openaicom_referred=true.

Pennsylvania Office of the Governor. "Shapiro Administration Sues Character.AI Alleging AI Chatbot Unlawfully Presented Itself as Licensed Medical Professional in Pennsylvania." Commonwealth of Pennsylvania. May 5, 2026. https://www.pa.gov/governor/newsroom/2026-press-releases/shapiro-administration-sues-character-ai-over-fake-medical-claim.

Pfefferkorn, Riana, et al. "AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims." Stanford Cyber Policy Center, 2025. https://doi.org/10.25740/mn692xc5736.

Portnoff, Rebecca, and Michael Simpson. *Safety by Design: Annual Progress Report*. Thorn, 2025. https://info.thorn.org/hubfs/Thorn_SafetyByDesign_AnnualProgressReport_April2024-April2025.pdf.

"Progress Update: Responsible AI and Child Sexual Abuse and Exploitation Online." Google, April 2025. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/ai_responsibility_and_csae_en.pdf.

Rad, Dana, and Gavril Rad. "Exploring the Psychological Implications of ChatGPT: A Qualitative Study." *Journal plus Education* 32, no. 1 (2023): 43–55. https://www.researchgate.net/publication/370543260_EXPLORING_THE_PSYCHOLOGICAL_IMPLICATIONS_OF_CHATGPTA_QUALITATIVE_STUDY.

Ravitz, Georgia C. "FDA's Digital Health Advisory Committee Considers Generative AI Therapy Chatbots for Depression." Orrick (blog). November 11, 2025. https://www.orrick.com/en/Insights/2025/11/FDAs-Digital-Health-Advisory-Committee-Considers-Generative-AI-Therapy-Chatbots-for-Depression.

Robb, Michael B., and Supreet Mann. *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Common Sense Media, 2025. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf.

Safeguarding and Improving Youth Mental Health in the Digital Era. Child Mind Institute, October 14, 2025. childmind.org/education/childrens-mental-health-report/youth-mental-health-in-the-digital-era/.

"Safety & Responsibility." OpenAI. Accessed December 10, 2025. openai.com/safety/.

"Social AI Companions." Common Sense Media, July 2025. www.commonsensemedia.org/ai-ratings/social-ai-companions?gate=riskassessment#section-5.

Social Media Victims Law Center. "AI Chatbot Companions' Impact on Children and Teens." Social Media Victims Law Center (blog). February 24, 2026. https://socialmediavictims.org/blog/ai-chatbot-companions-impact-children-teens/.

Stanja, Judith, Jessica Rose Meier, and Johannes Krugel. "Children's and Adolescents' Anthropomorphic Conceptions of Social Robots and Chatbots–A Systematic Literature Review." *In Proceedings of the 25th Koli Calling International Conference on Computing Education Research*, pp. 1-10. 2025.

"Strengthening ChatGPT's Responses in Sensitive Conversations." OpenAI (blog), October 27, 2025. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/.

"Taking Bold Steps to Keep Teen Users Safe on Character.AI." Character.AI, October 29, 2025. blog.character.ai/u18-chat-announcement/.

Thiel, David. "Investigation Finds AI Image Generation Models Trained on Child Abuse." Stanford Cyber Policy Center, December 20, 2023. cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

"Usage Policy." Anthropic. Accessed December 10, 2025. www.anthropic.com/legal/aup.

van der Goot, Margot J., et al. "Understanding Users' Responses to Disclosed vs. Undisclosed Customer Service Chatbots: A Mixed Methods Study." *AI & Society* 39 (2024): 2947–2960. https://doi.org/10.1007/s00146-023-01818-7.

Wang, Echo. "TikTok Hits 1 Billion Monthly Active Users Globally - Company." *Reuters*, September 27, 2021. www.reuters.com/technology/tiktok-hits-1-billion-monthly-active-users-globally-company-2021-09-27/.

Warner, Mark R. "Warner Introduces Bipartisan Bill Protecting Children from AI Chatbots with Parents, Colleagues." October 28, 2025. https://www.warner.senate.gov/public/index.cfm/2025/10/hawley-introduces-bipartisan-bill-protecting-children-from-ai-chatbots-with-parents-colleagues.

Wei, Marlynn. "The Emerging Problem of AI Psychosis." Psychology Today, November 27, 2025. www.psychologytoday.com/us/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis.