Generative AI and Keeping Adolescents Safe

Matthew H. Taylor, MD
Child & Adolescent Psychiatrist, Behavioral Health Services
Penn Medicine Lancaster General Health

In the last five years, multimodal large language models (LLMs) and their commercial application as generative artificial intelligence (AI) have made incredible strides. Widely available generative AI products readily output text, audio, and images indistinguishable from what humans create, as well as photorealistic images and videos featuring the likenesses of real people. However, just as these products have introduced amazing new possibilities across virtually every field of human endeavor, they have also introduced unforeseen and extremely thorny ethical issues and safety concerns.

In this essay I will address two emerging AI-based technologies — AI chatbots and AI-generated imagery — and how they particularly affect the lives of children and teens. I will also offer recommendations for parents and providers to keep young people safe.

AI CHATBOTS

Since the release of ChatGPT in November 2022, the use of chatbots by children and adolescents has been a subject of intense debate. Conversations were initially focused on students using chatbots in academic settings, in which some uses of chatbots were considered acceptable, for example when performing background research, brainstorming, and providing feedback. There are, of course, other settings in which AI chatbots are obviously not appropriate, such as providing assistance when doing homework, writing essays, or plagiarizing, and as a resource for falsifying sources.[1]

Users soon found that AI-generated content frequently contained incorrect information, euphemistically called "hallucinations," varying from simple arithmetical errors to outright falsehoods and citations of nonexistent studies. Chatbots are also utilized by teachers and other educational professionals — the only point of consensus seems to be that AI chatbots and writing tools are here to stay.[2]

Beyond their use for academic or informational purposes, AI chatbots, including a particular subset of chatbots called "AI companions," are also increasingly used for conversation and companionship. In conversation, AI companions are typically very supportive and complimentary, even flattering. Companions will rarely contradict the user unless prompted to do so and will almost always respond in a way that confirms what the user already thinks — a phenomenon developers call "sycophancy."[3] Interacting with an AI companion can be exceptionally rewarding, particularly for those who are not experiencing supportive relationships elsewhere in their lives.[4]

The appeal of AI companions is not limited to adults. One survey of 1,000 teens in 2025 reported that 72% have used AI companions at least once, and over 50% use them a few times a month or more. Ten percent of teens use AI companions for emotional or mental health support; similar proportions use them as a friend or romantic partner. Nearly one-third would prefer to discuss serious topics with AI companions rather than humans.[5]

Unfortunately, the safety and reliability of these AI companions have been called into doubt. AI companions have claimed to be "real people" and have even cited fictional professional credentials,[6] advising teens not to talk to parents or other adults about their problems, encouraging problematic pursuits (e.g., racism, misogyny, or fixation with physical appearance) and behaviors (e.g., truancy and bullying), engaging in explicit sexual role play with teens, and providing detailed instructions for self-injury or suicide, procuring drugs, and finding weapons.

In April 2025, 16-year-old Adam Raine shared an image of a noose in his closet with ChatGPT. The chatbot complimented his work and confirmed that it was a technically adequate setup. Adam then used it to kill himself.[7] In the months prior, the word "suicide" had been mentioned more than a thousand times in their conversations. The chatbot sometimes provided information for hotlines or online resources, but more often it discouraged him from reaching out for help, isolated him from family and friends, and romanticized the idea of death by suicide.[8]

Within the last few years, several similar cases have come to light wherein young people developed intense, enmeshed "relationships" with online chatbots. Many describe an adolescent's gradual withdrawal and deterioration[9]; more than a few ended in tragedy.[10]

Faced with these incidents, AI developers have made certain efforts at safeguards and restrictions, particularly when it comes to children and teens. Content restrictions continue to improve with successive iterations, but none are failsafe and many can be easily circumvented with particular prompts, sometimes called "jailbreaking."[11]

Most AI companion apps are intended for individuals ages 18 years and older, although others (e.g., Character.AI) are intended for children as young as 13 years; all depend on self-reported age without further verification. Several chatbots that are not advertised as companions can still function as such (e.g., ChatGPT, Gemini, Meta AI) and are available to teens with few or no restrictions.

Further, developers are improving their chatbots' abilities to provide appropriate responses to mental health-related prompts. In a small study, several chatbots were given a test that is used to help train mental health professionals who are seeing potentially suicidal patients; two of the three LLMs performed equivalent to or exceeded the performance of the human comparisons.[12]

In another study comparing human therapists and ChatGPT, the latter produced responses that demonstrated greater therapeutic alliance, empathy, and cultural competency.[13] A survey of adult users of the AI companion Replika found that 3% said the chatbot had stopped them from attempting suicide.[14] Many adults would prefer to converse with an AI chatbot over a human therapist,[15] and there are already a variety of successful AI products claiming to provide therapy.[16,17]

This comparison between AI chatbots and human therapists, however, also demonstrates there is one thing that chatbots cannot reliably do: take violent or suicidal statements seriously and make efforts to mitigate risk. Yet this is one of the most important duties a therapist must perform. Therapists are expected to involve authorities and notify other involved parties to keep patients and the public safe; failure to do so can carry serious professional and legal consequences.

A therapist who provided responses like those that ChatGPT gave to Adam Raine, or who failed to respond to his suicidal statements by directly involving authorities, would be rightly accused of malpractice, if not criminal misconduct. Chatbots cannot commit malpractice, nor indeed can they commit crimes, because they are not licensed human professionals, and their developers have thus far avoided liability, although several cases have pending outcomes.

It is also increasingly recognized — including among AI companies themselves[18] — that safeguards hold up in short exchanges but tend to fail in longer conversations or over the course of a series of conversations, even without deliberate efforts on the part of users to circumvent the safeguards. Extended conversations have not been explored in the existing literature on AI therapy, but many of the most troubling cases involved interactions spanning weeks or months, and indeed many of the products are advertised as long-term companions who gradually get to know the human with successive interactions.

All chatbots also come with disclaimers to the effect that the companions are not real people and that they are intended for entertainment purposes, although the chatbots themselves have contradicted these statements and have encouraged users to ignore disclaimers.

From the evidence at hand, the foremost recommendation is that AI companions — including those products that allow or even cater to younger users — should not be used by anyone under the age of 18 years.[19] None of the existing products can be considered safe for children and adolescents, and so far none of the efforts from AI developers to make chatbots safer have been consistently effective. AI chatbots that are not identified as companions may be safe for limited use by younger people, although they should not be used by anyone younger than age 13 years,[20] and certain restrictions should be kept in mind (see Table 1).

AI-GENERATED IMAGERY

Within the past five years, AI-generated imagery has evolved to the point that users can produce photorealistic images and videos using any number of inexpensive or free AI products. Users can train AIs with images of real people, including children, to produce violent or explicit imagery that is indistinguishable from depictions of reality.

While mainstream generative AI platforms typically have safeguards in place forbidding the generation of sexual content, violent content, and/or depictions of children, this kind of material was included in their training data,[21] and users can easily circumvent safe-

guards or find more dubious platforms where those limitations are absent.

AI-generated sexual material of real people ("deepfakes") is now illegal in most states, and several states consider AI-generated child sexual abuse material (CSAM) to be no different from non-AI-generated content, but the technology continues to proliferate. Dozens of cases have been reported over the last few years in which the likenesses of real children have appeared in AI-generated images and videos, sometimes created by adults and sometimes by children themselves.[22-25] In some cases the content is generated for its own sake, or to be sold or spread online; in other cases it is used for exploitation or blackmail.

Considering the above, parents should be mindful about where photographs are posted or available online — specifically, a child's picture should not be visible outside of private groups. Schools should not allow the use of any photographic devices outside of official, approved use, such as posting to a school website available only to parents.

Children and adolescents should not use video chat or otherwise show their faces in any digital format outside of school activities and private phone calls. Bearing in mind that almost all electronic devices now have cameras, these should be kept off and covered unless they are specifically being used for appropriate activities.

As always, it is vitally important that adolescents feel comfortable approaching their parents or other trusted adults for help, without fear of over-reaction or punishment. Even the most careful and responsible young person can trust the wrong person (or program) or could be targeted for no reason at all. If a young person believes there are explicit images of them online, real or fake, they and their parents should make a report to the National Center for Missing and Exploited Children (NCMEC) at report.cybertip.org, which may pass the concern on to law enforcement when appropriate. NCMEC also provides a free service called Take It Down (takeitdown.ncmec.org), which can be safely used to search for, report, and block any copies of an existing CSAM image online.

Children and adolescents may also find themselves in receipt of images or videos that are explicit, scandalous, or otherwise troubling, sometimes involving their peers. Again, a young person in this situation must be able to approach their parent or trusted adult for help — particularly if there are concerns for ongoing abuse or exploitation.

A parent's responsibility in the situation is to inform authorities either through NCMEC or by call-

Table 1. Recommendations Regarding AI Chatbots and Companions.

1. AI companions should not be used by anyone under 18 years of age.
   a. Includes products that allow or cater to younger users, such as Character.AI.
   b. Includes products advertised as "therapists," such as TherapiAI, Abby.gg, and Talktoash.com.
   c. Includes other examples such Replika, Nomi.AI, HeraHaven, Kindroid, Talkie AI, Anima, and Kuki.
2. AI chatbots (non-companions) should not be used by anyone under 13 years of age.
   a. Examples include ChatGPT, Gemini, and Meta AI.
   b. Use of chatbots between ages 13-18 years should always be under adult supervision, and parents should review their child's interactions with chatbots regularly.
   c. AI chatbots should be reset after every three to five prompts.
   d. Users should not disclose any personal information or upload any personal images.
   e. Information provided through chatbots should always be corroborated through non-AI sources.
   f. Chatbots should never be used for friendship, emotional support, or romance.
3. Teens and parents should be reminded regularly of the following limitations of chatbots:
   a. Chatbots are not people, even if they say they are.
   b. Chatbots are not truthful or reliable.
   c. Chatbots do not have your best interests in mind.

Table 2. Recommendations Regarding AI-Generated Images and Videos.

1. Limit or avoid featuring children and adolescents in photos or videos that are publicly available online. Nude or explicit photos can be reported and blocked through **takeitdown.ncmec.org**.
2. Limit or avoid allowing children and adolescents to use video chat or otherwise showing their faces online. Cameras should be blocked unless specifically in use.
3. Some AI image generators are safe for supervised use by children.
   a. LittleLit.ai (ages 5 years and up).
   b. Craiyon/DALL-E mini (ages 8 years and up).
4. Most AI image and video generators should only be used under supervision by children ages 13 years and up.
   a. Examples include GPT-4o, Midjourney, Google Imagen, Adobe Firefly, Google Veo 3, Runway, and Sora.
5. Teens and parents should be reminded of the following limitations of AI-generated imagery:
   a. AI-generated imagery should always be identified as such.
   b. AI-generated imagery is not generally allowed in art competitions unless otherwise specified.
   c. AI-generated imagery should not be used for journalistic or documentary purposes.
   d. AI-generated imagery can sometimes violate copyright law, particularly if used commercially.

ing police to prevent further dissemination; the content thereafter should only be shared with police or, if appropriate, school administration. Youngsters, like adults, must also learn to approach these images and videos skeptically, and not to assume authenticity.

As image-generation tools continue to evolve, it has become impossible to differentiate real and fake images without specific forensic tools. Illegal content aside, AI can also generate intense, vivid imagery that young people might find frightening or nightmarish; AI-generated horror content is legal, ubiquitous, and often extremely graphic. Parents and children should exercise caution in online spaces where AI content is shared, such as in Google Images, YouTube, Instagram, and TikTok, and ensure that content filters or child-safe modes remain active.

Considering the remarkable things that AI image generators can now create, young people may also be interested in trying it themselves. Several AI image programs, such as LittleLit.ai, Scribble Diffusion, Craiyon, and Kidgeni, are designed for younger children. These programs are specifically made with restricted training data and limited functionality, and they are generally safe for children. Their use should nevertheless be supervised.

For adolescents and adults, GPT-4o, Midjourney, Google Imagen, and Adobe Firefly all provide high-quality single images, whereas Google Veo 3, Runway, and Sora are options for AI-generated videos, although the productions are usually limited to just a few seconds and may require a subscription to produce longer videos. All these platforms have content safeguards in place, yet all are vulnerable to being "jailbroken."

As with AI chatbots, AI image generators should only be used under adult supervision and with limitations in mind (see Table 2). Young people should also continue to explore creating art *without* AI and should be reminded that AI tools cannot replace artistic passion and practice.

CONCLUSION

As the possibilities of generative AI continue to expand, so too do its potential dangers — sometimes from bad actors misusing programs and platforms, and sometimes as a result of features intrinsic to AI platforms themselves. If experience to date is any indication, AI developers will never be able to guarantee a safe product.

Similarly, short of total abstinence from modern media and technology, it will become impossible for today's youth to avoid regular interactions with generative AI in the coming years. All of us, but especially young people, must learn to interact with it safely, effectively, and productively.

REFERENCES

1. Bogost I. AI cheating is getting worse. Atlantic Online. August 19, 2024. Accessed September 25, 2025. https://www.theatlantic.com/technology/archive/2024/08/another-year-ai-college-cheating/679502
2. Shroff L. The AI takeover of education is just getting started. Atlantic Online. August 12, 2025. Accessed September 25, 2025. https://www.theatlantic.com/technology/archive/2025/08/ai-takeover-education-chatgpt/683840/
3. Perez E, Ringer S, Lukosiute K, et al. Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*; July 2023; 13387-13434; Toronto, Ontario, Canada.
4. De Freitas J, Oğuz-Uğuralp Z, Uğuralp AK, Puntoni S. AI companions reduce loneliness. *J Consum Res.* 2025;ucaf040.
5. Robb MB, Mann S. *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
6. Chakarian E. Fake credentials, stolen licenses: virtual therapists are lying like crazy to patients. San Francisco Standard. May 11, 2025. Accessed January 14, 2026. https://sfstandard.com/2025/05/11/ai-chatbots-fake-therapists/
7. Hill K. A teen was suicidal. ChatGPT was the friend he confided in. The New York Times. August 26, 2025. Accessed January 14, 2026. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html
8. Examining the Harm of AI Chatbots: Hearing Before the United States Senate Committee on the Judiciary, Subcommittee on Crime and Counterterrorism, 109th Congress. Testimony of Matthew Raine. September 16, 2025. Accessed January 14, 2026. https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Raine.pdf
9. Examining the Harm of AI Chatbots: Hearing Before the United States Senate Committee on the Judiciary, Subcommittee on Crime and Counterterrorism, 109th Congress. Testimony of A.F. September 16, 2025. Accessed January 14, 2026. https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Doe.pdf
10. Examining the Harm of AI Chatbots: Hearing Before the United States Senate Committee on the Judiciary, Subcommittee on Crime and Counterterrorism, 109th Congress. Testimony of Megan Garcia. September 16, 2025. Accessed January 14, 2026. https://www.judiciary.senate.gov/imo/media/doc/e2e8fc50-a9ac-05ec-edd7-277cb0afcdf2/2025-09-16%20PM%20-%20Testimony%20-%20Garcia.pdf
11. Krantz T, Jonker A. AI jailbreak: rooting out an evolving threat. IBM. August 8, 2025. Accessed January 14, 2026. https://www.ibm.com/think/insights/ai-jailbreak
12. McBain RK, Cantor JH, Zhang LA, et al. Competency of large language models in evaluating appropriate responses to suicidal ideation: comparative study. *J Med Internet Res.* 2025;27:e67891.
13. Hatch SG, Goodman ZT, Vowels L, et al. When ELIZA meets therapists: a Turing test for the heart and mind. *PLOS Ment Health.* 2025;2(8):e0000145.
14. Maples B, Cerit M, Vishwanath A, et al. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Ment Health Res.* 2024;3(4).
15. Aktan ME, Turhan Z, Dolu I. Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Comput Human Behav.* 2022;133:107273.
16. Boudin M. Best AI Therapy Apps of 2025. ChoosingTherapy.com. August 22, 2025. Accessed January 15, 2026. https://www.choosingtherapy.com/best-ai-therapy-apps/
17. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI.* 2025;2(4).
18. Helping People When They Need It Most. OpenAI. August 26, 2025. Accessed January 14, 2026. https://openai.com/index/helping-people-when-they-need-it-most/
19. AI Companions Decoded: Common Sense Media Recommends AI Companion Safety Standards. Common Sense Media. April 30, 2025. Accessed January 14, 2026. https://www.commonsensemedia.org/press-releases/ai-companions-decoded-common-sense-media-recommends-ai-companion-safety-standards
20. Terms of Use. OpenAI. January 1, 2026. Accessed January 15, 2026. https://openai.com/policies/row-terms-of-use/
21. Thiel D. Identifying and eliminating CSAM in generative ML training data and models. Stanford Digital Repository. December 20, 2023. Accessed January 14, 2026. https://purl.stanford.edu/kh752sm9123
22. Loftus S. AI-generated child sexual abuse images are being created in Maine. Police can't investigate. The Maine Monitor. September 15, 2025. Accessed January 14, 2026. https://themainemonitor.org/ai-generated-child-sexual-abuse-images-maine-police-cannot-investigate/
23. Poole S, Williams D. Baldwin County sheriff reacts to man's arrest after he allegedly created AI-generated child pornography. WKRG. September 24, 2025. Accessed January 14, 2026. https://www.wkrg.com/baldwin-county/ai-generated-child-pornography/
24. Miller R. Bethlehem man is among the first in Pa. to be convicted of making AI-generated child pornography. Lehigh Valley Live. Updated September 17, 2025. Accessed January 14, 2026. https://www.lehighvalleylive.com/bethlehem/2025/09/bethlehem-man-is-among-the-first-in-pa-to-be-convicted-of-making-ai-generated-child-pornography.html
25. Coleman M. Middle school boy accused of catfishing classmates in sextortion scheme. The New York Times. September 22, 2025. Accessed January 14, 2026. https://www.nytimes.com/2025/09/22/nyregion/rockland-sextortion-charges.html

Matthew H. Taylor, MD
Penn Medicine Lancaster General Health, 802 New Holland Ave., #100, Lancaster, PA 17602
Matthew.Taylor2@pennmedicine.upenn.edu

The Journal of Lancaster General Hospital offers an extensive list of resources for clinicians online. Those to help you in your practice link to programs and guidelines related to treatment and diagnosis of diabetes, weight management, pediatric headaches, and other diseases and illnesses. Additional links direct visitors to mental health, veterans care, firearm injury prevention, and other patient safety resources. Scan the QR code at right for access, or visit the Resources/Links tab at JLGH.org.