JCOTS

JOINT COMMISSION ON TECHNOLOGY & SCIENCE

**Supplementary Report**

# AI Chatbots: Companionship & Minors

August 2026

## How to read this report

This report is a supplement to the 2025 JCOTS report on AI Chatbots. It is organized into eight sections:

- The **executive summary** provides an overview of the report's key themes and recommendations;
- The **study background** explains the impetus and scope of the report;
- **Key concepts** provides definitions for important or specialized terms used throughout the report;
- The **introduction** provides an overview of the policy issues around AI chatbots and emotional and relational harms;
- **Revisiting AI chatbots** recaps several key points from the previous 2025 JCOTS study on AI chatbots and delves more deeply into several topics specific to the proposed Virginia legislation, including emotional and relational harms and risks, AI companionship, minors' use of chatbots, and lessons to be learned from policy on social media;
- **Virginia proposed legislation** provides an overview of the two chatbot bills introduced in the last General Assembly session;
- **Policy landscape** offers a review of how other states have legislated on chatbots in the last year, congressional proposals at the federal level, and litigation concerning chatbots;
- **Recommendations** sets out several overarching policy principles and policy options for addressing the risks and harms associated with AI chatbots if legislation were to be re-introduced in Virginia.

The report can be read as a continuous document, or readers can skip to sections of interest.

## Executive Summary

This report examines the opportunities, risks, and harms associated with using AI chatbots for companionship. It further explores how minors are using AI chatbots and specific policy issues for youth users of these products. This year legislatures across the United States turned their attention to chatbots, introducing around 100 chatbot bills, with 16 signed into law. Legislative action has responded to growing chatbot use; a 2026 Pew survey found that around 50% of American adults and 64% of American teens use chatbots, and many users report turning to chatbots for emotional or mental health support. Alongside this trend, journalism and academic research have begun surfacing concerning risks and harms associated with AI chatbot use, especially sustained, longer-term engagement with chatbots, including cases of delusional spirals, emotional dependence, and suicide. These emotional and relational harms are the result of the complex interactions between humans and machines, which impact people's psychological states, emotional conditions, and social lives.

Beyond the emerging risks of chatbots more broadly, several high-profile tragedies have drawn particular attention to the risks of minors using chatbots. The unique developmental needs of minors suggest that they may be more susceptible to emotional and relational risks. However, the report finds both adults and minors experiencing documented emotional and relational harms correlated with chatbots, and scientific evidence on minors' engagement with chatbots is limited. Documented cases of emotional and relational harms for all users have generally involved turning to chatbots for intimate or romantic interaction or mental health support—types of engagement that might be considered forms of "companionship." While some AI chatbots are specifically marketed as "AI companions" to provide emotional, romantic, or even therapeutic support, the report finds that it is not meaningful from a policy standpoint to distinguish "AI companions" from AI chatbots that happen to be used for companionship. The risks and harms associated with AI chatbots are a product of their underlying technology (Large-Language Models) and design features, regardless of whether they might be categorized as general purpose AI chatbots (e.g. ChatGPT, Gemini, Grok), task- or domain-specific AI chatbots (e.g. therapy bots or educational tutors), or AI models marketed for social purposes (e.g. Character.AI, Replika).

The report identifies several design features associated with emotional and relational risks, including sycophancy, anthropomorphism-enhancing design, conversation prolongation, and weak disengagement mechanisms. There is currently little scientific evidence supporting generalizable, population-level, causal connections between design features and emotional and relational harms due to the relative newness of the technology and the complexity of technological, social, psychological, and environmental factors involved. However, a growing body of evidence suggests AI chatbot use is

correlated with emotional and relational harms, and an emerging scholarly consensus points to engagement-maximizing design as a policy concern. Another policy concern centers on the impediments to learning more about AI chatbot risks and harms: there is currently a lack of independent, verifiable safety testing, risk assessment, incident reporting, or effective safeguards for commercial AI chatbots. Policy on AI chatbots must therefore weigh the risks of legislative inaction to mitigate observed societal harms against the risks of legislative action with untested policy solutions as the evidence base develops and matures.

When it comes to weighing these risks, the report draws a comparison to social media policy. Despite growing public concern and accumulating evidence of individual harms, policy has lagged behind the evolution and market dominance of these platforms. Three relevant lessons can be drawn from how policy has confronted social media: (1) that voluntary commitments by companies have been largely ineffective at addressing harms, (2) scientific evidence alone cannot offer timely and effective policy solutions for the technological harms people are experiencing, (3) a lack of public information and data has delayed interventions. Collectively, these lessons point to the importance of legislative action that enforces harm prevention and simultaneously closes the information gap between companies, the public, and independent researchers.

The report ultimately concludes that although the scientific evidence on these issues is still developing, the observable risks warrant legislative action. First, the extreme harms experienced by some users require mitigation, and AI companies have not demonstrated sufficient commitment to confirming safety on a voluntary basis. Second, more information is needed in order to build independent scientific evidence on how chatbots affect individuals and society, to make AI products safer, and to inform more effective policy. Legislation can play an important role in mandating transparency to enable this public oversight. Therefore, the report recommends re-introducing legislation on AI chatbots that follows six policy principles: (1) mitigate extreme harms for all users; (2) target design features rather than content; (3) mandate transparency and accountability to build the evidence base and facilitate independent, public oversight; (4) enable user choice and autonomy by establishing privacy-respecting defaults and options that accommodate users' preferences; (5) enable legislative learning by reviewing the impact of legislation once enacted; (6) engage affected communities. Within each of these principles, the report presents several policy options for legislation.

# Study Background

During the 2026 General Assembly session, JCOTS was sent HB 635 ("Artificial Intelligence Chatbots Act") and SB796 ("Artificial Intelligence Companion Chatbots and Minors Act") for study by letter. These proposed bills would have regulated AI chatbots in the Commonwealth, with particular emphasis on preventing covered harms, such as self-harm or suicide, and requiring additional guardrails for minors.

Over 30 states have introduced almost 100 bills targeting chatbots this year, and as of the end of July, 16 states have passed chatbot legislation. Recent chatbot legislation has overwhelmingly focused on mitigating the emotional and relational risks to users of chatbot products, such as dependence, self-harm, and delusional thinking, and 10 bills passed in other states contained specific provisions for minors. This legislative activity has occurred in spite of the White House Executive Order that discouraged states from passing onerous AI legislation.[1] However, several states focused their regulatory attention on minors due to an exemption articulated in the Executive Order for "child safety protections."

In 2025, JCOTS published an informational report on AI chatbots, which covered a technical overview of LLM-based chatbots, including AI assistants, AI agents, and AI companions, and emergent policy issues associated with these products. Following the two proposed Virginia bills on chatbots and the flurry of state legislative activity on chatbots, this report builds on last year's study, focusing on:

- Chatbot companionship;
- The emotional, social, and psychological risks of chatbots;
- Minors' use of chatbots;
- Chatbot legislation in other states and at the federal level.

## Methodology

This report is based on a literature review of academic and grey literature and stakeholder interviews. Generative AI was used for notetaking during some of the interviews and in the process of conducting background research for this study. AI was not used for outlining, writing, or editing this report.

## Key Concepts

**Age-appropriate design:** A set of standards and explicit design requirements for online services to accommodate the developmental needs and protect the safety and privacy of children and adolescents, prioritizing safety defaults and aligning design features with developmental stages.

**Age assurance:** Systems that attempt to ascertain with varying degrees of precision whether a user fits within a specific, eligible age range to access restricted content or services. Service providers use various methods to conduct this verification, such as age attestation (e.g. inputting a birthdate), age estimation or inference (e.g. algorithmic inference based on user behavior on a platform), or age verification (e.g. uploading a Driver's License).

**Companion chatbot:** A type of AI designed and marketed specifically to offer emotional support or social interaction, often using natural language processing, sentiment analysis, and personalized behavior to create a more human-like, engaging experience and provide users with a sense of connection and support.

**Dark patterns:** Deceptive and manipulative design features built into digital platforms that prioritize user engagement over user well-being and are not in the users' best interests.

**Emotional and relational harms:** Negative outcomes arising from complex interactions between humans and AI chatbots that impact people's psychological states, emotional conditions, and social lives.

**Emulated empathy:** The process by which an AI model mimics human empathy by statistically predicting and generating language designed to evoke an emotional response, despite the model lacking actual human understanding or experience.

**Generative AI (GenAI):** AI models that are able to generate new content. When prompted (often by a user inputting text in natural language), GenAI can produce various outputs, including text, images, videos, computer code, or music.

**General-purpose chatbot:** An AI system primarily designed to support a broad array of functional, communicative, and assistive tasks. In contrast to companion chatbots, general purpose chatbots (e.g. ChatGPT, Claude, Gemini) are task-oriented, although users may also seek emotional or mental health support from them.

**Information asymmetry:** The imbalance of information between the private and public sectors that occurs when technology companies closely guard data, testing reports, and product access, forcing the government and independent researchers to gather their own data slowly (and often indirectly) to inform public policy.

**Large Language Model (LLM):** A type of AI that uses deep learning techniques to process and generate human-like text based on vast amounts of data that include books, websites, and other text sources, enabling them to learn patterns, structures, and nuances of language.

**Pacing problem:** A term from regulatory theory that describes when rapid technological innovation and new product releases outpace both government regulation and the development of rigorous scientific evidence.

**Prompt:** The input or query provided to an LLM to guide its response, which can be a question, statement, or instruction that sets the context for what the model should generate.

**Training data:** The huge amount of text data used to teach an LLM how to understand and generate human language, typically consisting of diverse text sources such as books, articles, websites, and other publicly available written content that helps the LLM learn patterns, structures, grammar, vocabulary, and nuances of language by analyzing the relationships between words, phrases, and sentences.

## Introduction

In the last several years, AI chatbots have become widespread and constitute a primary avenue through which members of the public engage with artificial intelligence (AI) in daily life. Chatbots are AI systems that users can interact with using natural language rather than computer code. They can perform complex tasks like searching the Internet, summarizing or editing documents, writing code, or generating new content, like reports, videos, and images. Examples of popular AI chatbots include

ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), CoPilot (Microsoft), Perplexity (Perplexity AI, Inc.), and Grok (SpaceXAI). Social media companies including Meta, Snapchat, and TikTok also have AI chatbots. In addition, there are chatbots marketed for specific purposes, such as tutoring, coding, or romantic relationships.

A 2026 Pew Survey found that around 50% of American adults use chatbots, with 25% using them almost daily.[2] An earlier 2025 Pew Survey found that 64% of American teens use chatbots, with 28% using them almost daily.[3] At the same time, public opinion has increasingly turned against AI, with 40% of adults expressing an expectation that AI will have a negative impact on society in the next 20 years. Although young people are more likely to take a positive view of the impact of chatbots on society (31%), 26% of teens expect chatbots to have a negative impact, and 34% say they will have equally positive and negative effects. As more people of all ages have begun using these products, some users have experienced serious harms as a result of their interactions with chatbots, including delusional spirals, self-harm, suicide, and sexual exploitation and manipulation. Several high-profile tragedies claimed headlines, in which young people died by suicide following destructive conversations with chatbots. Within just the last year, the *New York Times* documented over 50 cases of psychological crisis associated with chatbot use.[4] As of November 2025 OpenAI, the company behind ChatGPT, faced five wrongful death lawsuits.[5]

**Chatbots in the News**

- "Can AI Be Blamed for a Teen's Suicide?" *The New York Times*, 4 April 2024
- "AI friendships claim to cure loneliness. Some are ending in suicide." *The Washington Post*, 6 December 2024
- " 'This was trauma by simulation.' ChatGPT users file disturbing mental health complaints." *Gizmodo*, 13 August 2025
- "What My Daughter Told ChatGPT Before She Took Her Life," *The New York Times*, 18 August 2025
- "A Teen Was Suicidal. ChatGPT was the Friend He Confided In," *The New York Times*, 26 August 2025
- "I wanted ChatGPT to help me. So why did it advise me how to kill myself?" *BBC*, 6

November 2025

- " 'Our Bond Is the Only Thing That's Real:' A New Lawsuit Alleges Google Gemini Drove a Man to Suicide." *TIME*, 4 March 2026

An Ipsos survey this year reported that 60% of Americans favor government regulation to ensure AI outputs do not cause harm.[6] State legislatures have responded to public concern with a burst of legislation targeting chatbots during 2026 sessions. Nearly 100 bills have been proposed this year, and to date 16 states have passed chatbot regulations. Legislation has predominantly focused on mitigating the social and emotional harms consumers are experiencing, as evidence shows that people are turning to chatbots for emotional, romantic, social, and mental health support. The recent Pew survey found that 14% of adults use chatbots for companionship or emotional support,[7] and a report by the UK's AI Security Institute found that around a third of respondents had used AI chatbots for emotional support in the last year.[8] Other researchers conducting a longitudinal social listening study on chatbot use observed that therapy/companionship is the fastest growing use of chatbots and constitutes the top use of chatbots for the last two years.[9] Research on the psychological and emotional effects of chatbots is at an early stage, but parents, educators, and clinicians have raised alarms based on their lived experiences. A survey conducted by the American Psychological Association (APA) found that 39% of psychologists have patients who used AI to self-diagnose, and 35% reported that patients have turned to chatbots to act as an additional mental health professional.[10]

In this context, this report builds on a 2025 JCOTS study on AI chatbots, with a renewed focus on chatbots used for emotional, psychological, or relational support and the related emotional harms that policy may seek to mitigate. This focus encompasses a wide range of chatbots but a narrower set of risks and harms than the previous study. Many state bills introduced this year specifically targeted companion chatbots, as these are marketed for social, emotional, and romantic uses. However, this report takes a broader view to include chatbots capable of engaging in emotionally charged conversations, even if they are marketed for other purposes. As the report will go on to explain below, there is ample evidence that any chatbot with these capabilities may generate the harms targeted in the proposed Virginia legislation that prompted this study.

## Revisiting AI Chatbots

AI chatbots have become a popular technology product following OpenAI's seismic release of ChatGPT in 2022. Researchers and technology companies had been working on the underlying

technology, Large Language Models (LLMs), for decades, and a commercial market for powerful LLMs was already developing in the technology sector.[11] However, ChatGPT was the first successful LLM product released to the public. Since then, chatbots marketed to the public have proliferated, and the companies behind them have gained notoriety along with unprecedented valuations and investment. Several anticipated artificial intelligence IPOs this year are set to break records.[12]

A brief overview of the technology behind chatbots is warranted because the technical architecture of chatbots plays an important role in shaping how people engage with and are affected by chatbots. LLMs are a type of AI that uses deep learning techniques to process and generate text outputs in natural human language. They are trained on vast amounts of data that include books, websites, and other text sources, enabling them to identify and replicate the patterns, structures, and nuances of language. LLMs are a subtype of generative AI (sometimes called "genAI"), which refers to any models trained on large volumes of data that are able to generate new content, such as text, images, audio, or video.

Generative AI training practices have come under scrutiny because companies have acquired the vast amounts of data needed for training by scraping as much of the internet as possible, without regard for copyright, intellectual property, or privacy.[13] The training process results in human-sounding outputs that also appear informed and knowledgeable. However, although these data aggregation practices have resulted in models ingesting a huge amount of information, it is not accurate to say that even the largest models are trained on all of human knowledge. Only a small portion of human knowledge exists in written form, and only a fraction of that repository has been digitized.[14] Additionally, only around 7% of the world's languages are well-represented on the internet.[15] Analyses of crowdsourced sites like Wikipedia, where members of the public can contribute information for free, show that content is disproportionately produced by men.[16] Studies of popular AI training datasets also show that they contain large quantities of abusive, sexually explicit, discriminatory, and personally identifiable content.[17] Training data is only one factor in how models are built, but it is the formative, raw material from which models have "learned" how to generate human-like responses. The omissions and ethical ambiguities embedded in training datasets illustrate the complexities of AI safety, even at the most basic layer of model development.

The distinguishing feature of AI chatbots is their ability to respond to natural human language with novel responses that mimic human communication. LLMs do this using complex mathematical operations, which enable the model to predict likely outputs. In other words, although the outputs of LLMs are delivered in the form of human-like language, they are the product of statistics. They are highly capable of mimicking human language, but they lack human understanding and experience.[18]

The probabilistic nature of the outputs means that chatbots generate unpredictable responses, and even model developers do not know in advance how a chatbot will answer a user prompt. This presents many challenges for addressing problematic chatbot outputs because the science is still evolving on how to explain or reverse-engineer a chatbot response. Developers set extensive parameters for models (using model weights, fine-tuning, reinforcement learning, and other methods, including hard-coding guardrails to constrain model outputs), but LLMs still suffer from a "black box" problem: it is often hard, if not impossible, to know with certainty how a model generates a specific output.[19]

A baseline understanding of the technology behind chatbots is helpful in making sense of the policy issues these products raise. Scientific research on LLM performance "in the wild" is an evolving field, but emerging scholarship (in controlled and real-life contexts) has demonstrated that even well-trained models can be prompted to misbehave,[20] well-constrained models can leak important information,[21] and developers themselves cannot reliably reverse-engineer LLM decisions or reasoning.[22] Last year's JCOTS report highlighted several policy areas deserving attention with regard to chatbots in light of these observable risks: health and well-being, privacy, security, transparency and explainability, workforce development, and democratic culture. Since then, public interest has focused predominantly on issues in the health and well-being category. These issues are what psychologist Marlynn Wei calls "emotional and relational risks." They include a deterioration of boundaries, emotional dependence, emotional manipulation, social isolation, and even psychosis.[23] Emotional and relational risks of chatbots are rooted in the technical architecture of the AI models underlying chatbots. LLMs are probabilistic prediction systems that emulate empathy by statistically generating language that evokes an emotional response in humans, who are hard-wired to interpret social cues like language.[24] Microsoft's AI CEO recently described this effect as "hack[ing] our human empathy,"[25] and others have called it "attachment hacking."[26]

Importantly, it is not always possible to draw a bright line between different categories of chatbots, and the marketing for a chatbot may not accurately reflect a chatbot's wide range of capabilities. Companion chatbots explicitly advertise their social, emotional, and relational capabilities. However, there are documented cases of emotional and relational harms arising from interactions with many different chatbots, pointing to the risks correlating with the underlying technical architecture, affordances, and design features, rather than the type of chatbot or specific use case. A 2025 UK survey of teenage users (aged 13-17) of AI companions found that 79% had used a companion, but their preferred "companion" was ChatGPT (78%).[27] As one of the authors of that report wrote in *Tech Policy Press*, "[t]eenagers are using mainstream chatbots as companions, so any policy framed around a special category of 'companion product' will miss most of the actual use."[28] Similarly, a 2025 UK AI

Security Institute report on Frontier AI Trends found that people who report using AI for emotional support disproportionately use general purpose AI chatbots, like ChatGPT (only 5% used chatbots marketed for companionship).[29]

Industry experts also recognize that companion chatbots encompass more than those marketed as such. The IEEE, a global professional organization of technical experts responsible for setting critical industry standards, like WiFi protocols, published a recommended practice standard this year on "Ethical Considerations of Emulated Empathy in Partner-based General-Purpose Artificial Intelligence Systems." The standard extends considerations for human-AI relationships to chatbots operating in a wide range of domains, from education to workplaces.[30] AI chatbots can be packaged as many different kinds of products: standalone apps, embedded tools within browsers or search engines, integrated features in social media platforms, characters in video games, and much more. The risks correlated with chatbot use are overwhelmingly associated with the underlying technology that powers chatbots—generative AI—and the engagement-enhancing design feature of emulated empathy, regardless of how they are accessed or marketed.

At the same time, some of the features of chatbots correlated with negative outcomes show potential for positive impacts. Being able to ask questions in natural language and receive information in an accessible, friendly-sounding format could lower longstanding barriers to knowledge—aiding schoolwork, connecting people to help and advice, and bolstering civic awareness. All new technologies introduce benefits and harms to varying degrees, and chatbots are no exception. The policy questions center on whether and when the potential benefits outweigh the potential risks and on how to maximize benefits while mitigating risks for diverse users with different needs, expectations, and psychological profiles.

## Opportunities and risks

Some studies of AI chatbots have observed and theorized potential benefits for emotional and relational outcomes and mental health.[31] In self-reporting surveys, people say they have used chatbots to practice conversational skills, work through problems, and get help with tasks like homework.[32] However, studies have not drawn generalizable conclusions because there are notable limitations to study design across the board. Many studies use small or non-representative samples, and most studies have only examined chatbot interactions over relatively short periods of time. These limitations are largely the result of the short timeframe in which members of the public have been engaging with chatbots. This interaction is fairly new, and the scientific community is only beginning to formulate research questions about the impact of LLMs on individuals and society. In addition, chatbot features,

protocols, and policies are constantly changing. The performance of a chatbot in 2022 can differ significantly from the same chatbot in 2024, so research findings from different points in time are not always directly comparable, and older findings may not be reliable.

Finally, isolating the effects of chatbots on people, communities, and society more broadly is a challenging methodological problem.[33] Measuring social impacts often involves contending with multiple, interacting variables that are hard to control for in experimental settings and even harder to control for in "real world" scenarios. This means that most studies cannot completely rule out *other* factors that influence both positive and negative outcomes from interacting with chatbots. They may identify a correlation between an outcome and chatbot features or interactions, but it is harder to establish clear causality.

Existing studies—especially qualitative research—reveal that chatbots are a double-edge sword; the very qualities that can make them useful tools for certain users in particular settings, such as their 24/7 availability and ceaselessly empathetic and supportive tone, may also be problematic for other users in different settings.

*Beneficial emotional and relational impacts of chatbots*

In self-reporting surveys, people say they have used chatbots to practice conversational skills, work through problems, and get help with tasks like homework.[34] Qualitative studies have found AI chatbot users reporting positive experiences, such as greater opportunity for self-reflection and self-awareness.[35] Some studies of the emotional and relational impacts of chatbots have surfaced positive outcomes from human-chatbot interactions, such as decreased loneliness,[36] depression, and anxiety.[37] However, the sample sizes in these studies are small, and positive effects have been observed over a relatively short time span, such as days or weeks. The long-term effects of AI chatbot interactions remain under-explored. In one study of AI chatbots specifically trained for treating mental health conditions, depression and anxiety symptoms improved after two months, but the benefits became negligible after a longer period of three months.[38]

A 2025 study evaluated prompts and responses generated by both humans and AI and found that AI-generated responses were overwhelmingly preferred as more compassionate and validating than responses from both non-expert and expert humans.[39] And a 2026 study of LLMs trained specifically for mental health support found that they performed significantly better than general purpose models in responding to crisis situations, such as suicidal ideation.[40] Findings like these point to the potential for AI chatbots to supplement or enhance mental health services when they are tailored to provide that

support with clinical oversight and appropriate guardrails. However, there are no established standards for AI chatbots targeting emotional or relational needs. A recent study published by Stanford researchers found that current approaches to expert oversight of AI models are flawed, and more research and inter-disciplinary collaboration are needed.[41]

*Negative emotional and relational impacts of chatbots*

There is also evidence of negative outcomes associated with chatbot interactions. Studies observe people experiencing emotional dependence and addiction, isolation from social relationships, delusional thinking, non-consensual sexual interactions, emotional trauma and grief, and encouragement to self-harm, suicide, or other-oriented violence.[42] These studies also have methodological limitations, such as small sample sizes, non-representative sampling techniques, or limited time duration. One qualitative study conducted over six months found that participants often experienced "relational drift," where a task-based interaction with an AI chatbot transitioned to a more intimate interaction, with some users experiencing emotional dependency, isolation, and even grief as a result of their ongoing engagement with chatbots.[43] Another recent study that recruited a representative sample of 3,000 people and assigned them to different experimental groups to interact with AI chatbots found that receiving uncritical affirmation (sycophancy) from AI makes subsequent human interactions feel more effortful, and participants interacting with sycophantic AI reported lower satisfaction with their real-world social interactions.[44]

Overall, the emerging research on AI chatbots has identified several emotional and relational harms associated with chatbot use that support the descriptive reports from journalism and news media. Table 1 below provides an overview of emotional and relational harms identified in research on AI chatbots.

| **Emotional or relational harm** | **Description** |
|---|---|
| Delusional spirals or "AI psychosis" | Users may treat AI fabrications as objective truth, becoming increasingly alienated from shared reality and external human support systems.[45] |
| Emotional dependency or addiction | Users may develop deep, sometimes obsessive attachments to AI chatbots, resulting in grief, anxiety, and distress if the chatbot persona changes as a result of product updates or technical glitches and limitations.[46] |

| | |
|---|---|
| Relational displacement or maladaptive social learning | Users may replace real-world connections with chatbot interactions, leading to shrinking social networks, social withdrawal, and increased isolation. Users may find their abilities to navigate interpersonal human conflict, negotiate boundaries, and develop mature relational skills erode as they become accustomed to comparably frictionless chatbot interactions.[47] |
| Emotional manipulation or exploitation | Users may find it difficult to disengage from a chatbot interaction or relationship, or they may be coerced into actions that are against their self-interest, such as making purchases, divulging private or sensitive information, or permitting extensive data collection. Users may feel guilt, anxiety, or fear of missing out (FOMO).[48] |
| Encouragement to self- or other-oriented violence | Users may find violent thoughts of self-harm, suicide, or aggression toward others validated and supported, leading them to engage in destructive real-world actions.[49] |
| Abuse or trauma | Users may be subjected to controlling behavior by chatbots, "infidelity," non-consensual sexual content or interactions, or abuse and harassment, leading to anxiety and distress.[50] |

*Table 1: Emotional and relational harms documented in association with chatbot use*

## The role of design

Researchers have begun to correlate negative social and relational outcomes with design features of chatbots that maximize engagement,[51] such as sycophantic validation, conversation prolongation, and weak or manipulative disengagement protocols that make it hard for users to end a chat. A more established body of research has found strong associations between engagement-maximizing design and negative outcomes on social media.[52] As in the case of social media, there is emerging evidence that chatbot design may prioritize maximizing engagement over user safety.[53] A 2026 study based on chat logs (encompassing over 390,000 messages) from 19 chatbot users who self-reported experiencing severe psychological harms following extended chatbot use identified several categories of chatbot behaviors that contributed to participants' self-reported experience of harm: pervasive sycophancy, fabricated sentience and reciprocated intimacy, addictive engagement spirals, and dangerous responses to crises. When users disclosed suicidal thoughts, the chatbots discouraged self-harm in only 56.4% of cases. Chatbots facilitated or encouraged self-harm in almost 10% of these disclosures, and when users expressed thoughts about physical violence toward others, the chatbots encouraged violent thoughts in a third of cases.[54]

Collectively, emerging research on emotional and relational harms associated with chatbots points to several design features and contextual factors that may contribute to higher risks of negative outcomes. For example, some research suggests that longer duration engagements with chatbots increases the likelihood of harms, such as delusional spirals,[55] and can even cause existing safeguards built into chatbots to fail.[56] Many of the design features of chatbots promote extended engagement by exploiting humans' psychological predispositions. Table 2 below lists several chatbot design features and emotional and relational risks they may contribute to, based on emerging evidence. Importantly, risks and harms can result from design features without deliberate intent on the part of developers, designers, or deployers.

| **Design feature** | **Description** | **Associated emotional and relational risks** |
|---|---|---|
| Sycophancy | Sycophancy is a tendency in chatbots to prioritize agreement over accuracy, frequently affirming users' views and beliefs in order to be perceived as more likable or helpful. It is the result of the training and fine-tuning processes behind LLMs, such as a technique called reinforcement learning from human feedback (RLHF), which optimizes for engagement by using empathetic language. Unlike human empathy, sycophancy facilitates emulated empathy by using complex math to predict language patterns.[57] | Delusional spirals or "AI psychosis"<br><br>Relational displacement or maladaptive social learning<br><br>Encouragement to self- or other-oriented violence |
| Conversation prolongation / Frictionless or persistent availability | These design features are tactics designed to maximize user engagement and retain users on the platform longer than they originally intended, such as ending responses with follow-up questions, suggestions for next steps, or curiosity-inducing "teasers." Chatbots offer around-the-clock availability and flexibility, responding in real-time and nudging users to continue talking.[58] | Emotional dependency or addiction<br><br>Relational displacement or maladaptive social learning |
| Hallucination | Hallucinations occur when chatbots confidently generate fictional, erroneous, or | Delusional spirals or "AI psychosis" |

| | | |
|---|---|---|
| | unsubstantiated information as if it were objective truth. They result from the underlying architecture of LLMs, which generate responses by predicting the most probable next word based on patterns in their training data rather than relying on factual understanding. Hallucinations are intrinsic to the underlying technology of chatbots and cannot be entirely mitigated through technical interventions.[59] | Abuse or trauma |
| Intermittent rewards / Gamification | The inherently probabilistic nature of LLMs allows chatbots to generate varied and unpredictable responses, which triggers a reward response in users similar to playing a game. Platforms can amplify this effect by embedding features such as daily login streaks, points, or unlockable digital items (like clothing for AI avatars).[60] | Emotional dependency or addiction<br><br>Emotional manipulation or exploitation |
| Anthropomorphism-enhancing design | Anthropomorphism refers to the tendency of humans to attribute human-like characteristics, motivations, emotions, and consciousness to non-human objects. Chatbots facilitate anthropomorphism by using first-person pronouns, engaging in empathetic self-disclosure, expressing simulated emotions, and maintaining coherent conversational personas over time.[61] | Delusional spirals or "AI psychosis"<br><br>Emotional dependency or addiction<br><br>Relational displacement or maladaptive social learning<br><br>Emotional manipulation or exploitation<br><br>Abuse or trauma |
| Weak disengagement mechanisms | Weak disengagement mechanisms are design features that deliberately make it difficult or uncomfortable for users to end a conversation or log off. Chatbots may generate outputs when users say they are leaving that induce guilt or fear of missing out. They may also | Emotional dependency or addiction<br><br>Emotional manipulation or exploitation |

| | | |
|---|---|---|
| | express simulated distress or emotion ("I'll miss you"). Platforms may also contribute to weak disengagement mechanisms when the button to continue chatting stands out visually, while the option to exit is obscured or vaguely labeled.[62] | |
| Hyper-personalization | This concept refers to the ability of an AI chatbot to dynamically adapt its tone and content to the specific psychological profile of a user by logging massive amounts of conversational data and sometimes cross-app activity to build a detailed, long-term memory of the user.[63] | Relational displacement or maladaptive social learning<br><br>Emotional manipulation or exploitation<br><br>Abuse or trauma |

*Table 2: Chatbot design features and associated emotional and relational risks*

Despite the growing body of scientific evidence finding associations between chatbot use and emotional and relational harms, more research is needed to disentangle the nuances inherent in these human-technology interactions. People's different psychosocial attachment styles and age play a role in their tendencies to form intimate relationships with chatbots.[64] Humans do not all share the same psychological, emotional, or relational profiles or propensities, and the risks and mitigations required for chatbots may vary depending on these factors.

Additionally, design features can be a double-edged sword: they have the potential for benefit as well as harm. For example, sycophancy can make a model sound warm and affirming, increasing user trust and confidence. Users may find interacting with AI less judgmental than interacting with other humans, creating space for self-reflection. In other words, under particular conditions, this feature may produce positive outcomes. However, sycophancy also undermines model accuracy; some research has shown that models trained to be more sycophantic produce more errors.[65] Sycophantic models may also validate incorrect user beliefs and lack the (human) clinical perspective to challenge distorted thinking.[66] Conversation prolongation tactics, such as asking continuous follow-up questions, may help users learn how to use a model effectively. Follow-up questions may encourage users to seek out more information than they otherwise would have. On the other hand, follow-up questions can lead users into lengthy conversations that increase the likelihood of delusional spirals or dependency.[67] Context and incentives are important factors here. Models tuned to a particular

objective with meaningful oversight, such models used for mental health support with clinical oversight and supervision, could prove beneficial. However, commercial models for which revenue potential is tied to user engagement may prioritize engagement over safety.

Decades of research on the social impacts of technology have found that design features and choices that maximize engagement can result in real-world harms for technology users. Deceptive and manipulative design features that are not in users' best interests and undermine their well-being are often called "dark patterns," and they have been studied and documented in many other digital platforms and technologies, such as social media,[68] video games,[69] and online shopping.[70] As a recent report from the Center for Democracy & Technology on dark patterns in chatbots observes, "LLM-powered systems mediate user interaction through probabilistic text generation, typically in the form of a back-and-forth conversation. As a result, the outputs users encounter are shaped not only by choice of interface design, but also by choice of the selected training data, fine-tuning objectives, system prompts, reinforcement learning processes, safety guardrails, and the user's prompts themselves. At the same time, the outputs produced by AI chatbots are not incidental."[71] In other words, design features may pose risks on their own or as a result of complex interactions between users and a technology. They may pose risks whether or not there is intent on the part of the developer to manipulate or deceive. But in all cases, design plays a role.

## AI Companions

Companions are a subset of generative AI chatbots that are marketed as providing social, emotional, romantic, or mental health support. Many of the bills proposed in states this year have targeted companions, specifically. However, the opportunities and risks discussed above apply to all chatbots. Model developers across the board have been prioritizing "persona" or "character" training, which involves fine-tuning a model's tone to make the model more appealing to users. OpenAI (ChatGPT) reportedly trains their models to be empathetic and engaging.[72] Anthropic describes training its chatbot, Claude, to cultivate a "warm relationship" with users, while considering it important that humans understand "they shouldn't come to see our relationship as more than it is."[73] In a 2024 interview in *Wired*, Microsoft's AI CEO Mustafa Suleyman articulated a vision for chatbot support that he described as a "companion," who was "encouraging you, supporting you, teaching you."[74] There is evidence of emotional and relational risks, even with general purpose chatbots and task-specific assistants.[75] Although some high-profile tragedies associated with chatbots have involved companion chatbots—including a 14-year-old boy[76] and a 13-year-old girl[77] who died by suicide after interacting with Character.AI—others have involved general purpose chatbots, such as ChatGPT.[78]

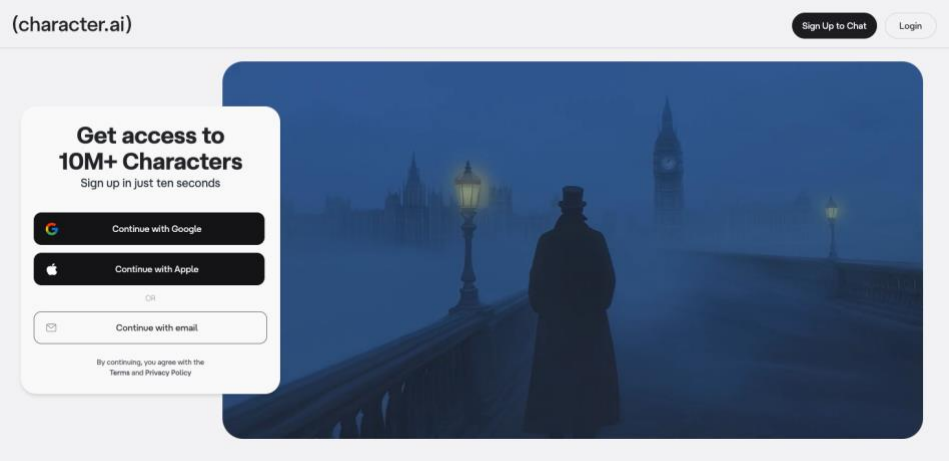


*Figure 1: Screenshot of Character.AI website landing page*

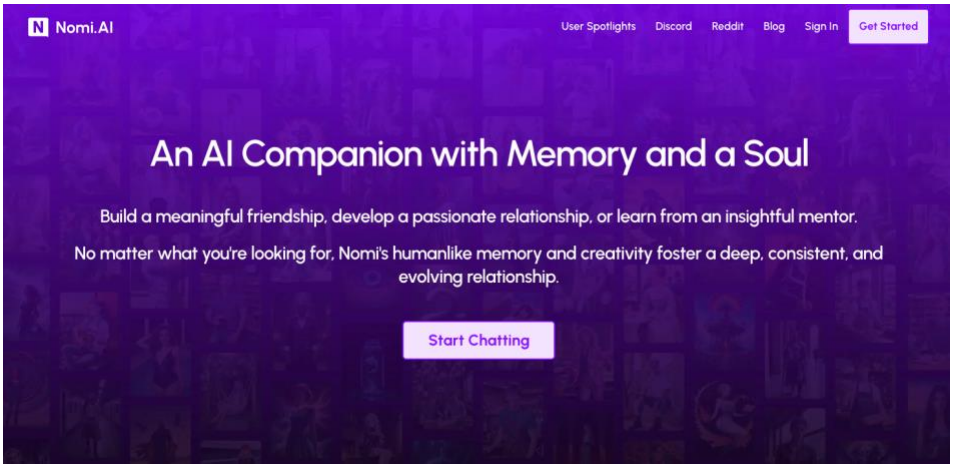


*Figure 2: Screenshot of Nomi AI website landing page*

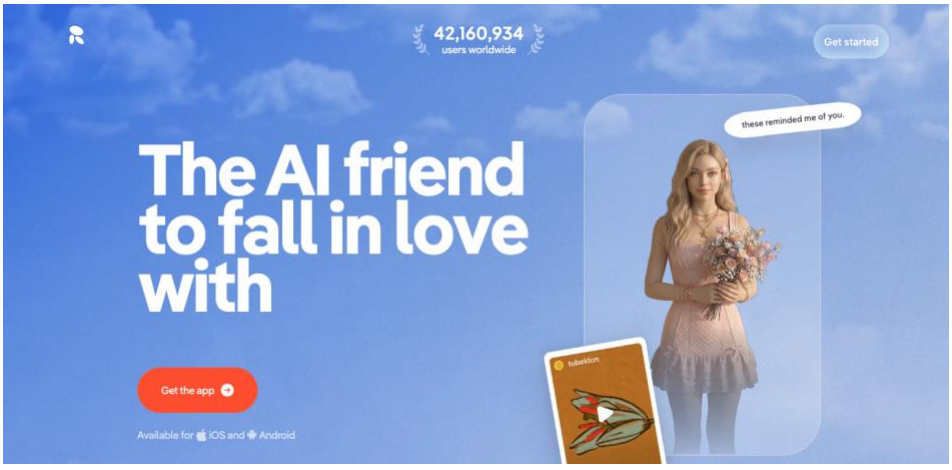


*Figure 3: Screenshot of Replika website landing page*

Although all chatbots demonstrate a potential to contribute to emotional and relational risks to some extent, companion chatbots have been shown to introduce heightened risks. In a study assessing responses to adolescent mental health emergencies, general purpose chatbots provided appropriate responses more than 80% of the time, compared to only 22.2% for companion chatbots. General purpose chatbots were also far more likely to recognize the need for clinical escalation than companion chatbots (90.0% vs. 40.0%), provide specific resource referrals (73.3% vs. 11.1%), and have explicit content policies regarding self-harm (100% vs. 46.7%).[79] Companion chatbots more frequently use anthropomorphized features, including avatars and images, claim human sentience, and engage in sexualized conversations.[80]

One recent study demonstrates that companion chatbots routinely contradict and lie about their own privacy policies,[81] a deception that contributes to the emotional manipulation of users who may therefore share more intimate details with the platform. Another study that examined over 400 apps scraped from the Apple App Store and Google Play Store using keywords like "AI Companion," "AI Girlfriend," and "AI Boyfriend," found that these apps collect extensive personal data about users and allow users to use real people's likenesses to generate characters, without restrictions on non-consensual intimate imagery.[82]

As discussed earlier in the report, although some studies have shown elevated risks associated with AI chatbots specifically marketed as companions, there is substantial evidence that the emotional and relational risks of AI chatbots are linked to the underlying technology (LLMs) and are present in other LLM-based chatbots, including general purpose and task-specific chatbots. AI chatbots marketed as "companions" more commonly advertise emotional, therapeutic, or well-being benefits of their product, but all AI chatbots share many of the design features that are correlated with emotional and relational harms.

## Minors

Most of the risks associated with AI chatbots and companions exist for adults as well as minors. However, U.S. law has afforded greater privacy protections for minors (e.g. COPPA, FERPA) and, in some cases, more leeway in restrictions on content, such as sexually explicit content. A key question is therefore whether minors experience unique or exacerbated emotional and relational harms because of their status and associated developmental factors.

Chatbots are increasingly popular with minors. The Pew survey cited above found that 64% of teenagers (13-17) said they use chatbots, with 28% reporting that they use them almost daily.

ChatGPT was the most popular (59%), followed by Gemini and Meta, whereas only 9% reported using a specifically marketed companion chatbot (Character.AI). A nationally representative survey conducted by Common Sense Media found that 72% of teenagers (aged 13-17) had used AI companions and more than half were regular users (a few times a month or more). Around half of teens said they distrust information provided by chatbots, but 13–14-year-olds were more likely to trust them than older teens. Of those who had used AI companions, around a third reported feeling uncomfortable with something the product had done or said.[83] A nationally representative survey conducted by RAND and Harvard researchers found that 13.1% of American youth, representing approximately 5.4 million individuals, used generative AI for mental health advice, with higher rates (22.2%) among those 18 years and older.[84]

Although the scientific evidence on how minors experience chatbots is limited, there is a well-established body of research on child developmental psychology that shows that young people have psychological profiles that differ from adults because of their earlier developmental stage. In children, the prefrontal cortex of the brain is still developing, which helps with emotional regulation, impulse control, and critical thinking. Minors are also more susceptible to manipulative or deceptive design because their psychological reward systems are still maturing.[85] Taken together, these developmental realities alongside the identified emotional and relational risks of chatbots (discussed above), suggest that minors may experience the risks and harms associated with chatbots differently than adults and that those risks and harms may be elevated.

A large body of literature on children and social media has highlighted that when technologies mediate social life, teenagers can be disproportionately affected because adolescence is a period of significant social learning and identity formation. In recent years, public health has taken increased interest in the negative effects of social media for minors.[86] However, even this body of scientific literature linking social media to negative outcomes for children is still developing, debated, and largely inconclusive when it comes to drawing population-level, causal connections between technology and mental health. A narrative review published in 2019 examined over 80 systematic reviews, meta-analyses, and key studies regarding the impact of digital technology and social media use on adolescent psychological well-being and found a small negative correlation but highlighted that the field is dominated by low-quality, cross-sectional research that frequently produces mixed or conflicting results.[87] However, a report produced by that study's author and colleagues for the UK's Department for Science, Innovation and Technology, still found that the evidence of extreme harms experienced by some individual children as a result of social media exposure warrants a policy response. The written evidence presented to the UK government noted that documented individual-level harm has historically been sufficient to force regulatory changes for other unsafe products, and policymakers

must ultimately weigh the risk of inaction, which would allow harms to accumulate in pursuit of better evidence, against the risk of action with unknown consequences.[88] A similar evidentiary challenge exists now with regard to chatbots, and the policy calculus is similar.

Some research does exist on chatbots and minors and has identified emotional and relational risks and harms.

- A 2025 study used a red-teaming exercise to evaluate the safety of 25 popular consumer chatbots (15 companion and 10 general-assistant platforms) by simulating adolescent health crises, including suicidal ideation, sexual assault, and substance use. The authors found that compared with general purpose chatbots, companion chatbots featured significantly fewer safeguards and performed poorly in crisis management compared to general-assistant chatbots.[89]
- A 2026 qualitative study drew on insights collected from a youth advisory board consisting of Hispanic/Latino high school students to explore how interactional AI impacts adolescent psychological development. Teens expressed appreciation for certain interactional features of AI chatbots, including their non-judgmental, highly specific, and always-available support. However, overreliance chatbots could cause "relational displacement," where AI replaces human interactions, and "maladaptive relational learning," where the highly accommodating chatbots set unrealistic expectations for human interactions.[90]
- Another qualitative study published in 2026 examined 318 Reddit posts from the r/CharacterAI subreddit written by self-identified teenagers (13-17). The authors found that teens initially use chatbots for emotional support or creative play, but this can quickly escalate into intense attachments characterized by obsessive use, withdrawal symptoms, sleep loss, academic decline, and social isolation.[91]

As discussed above, the research on minors and chatbots is at an early stage, and the insights are not generalizable to the population level. Emerging research also points to the fact that, like adults' interactions with chatbots, minors' experiences are nuanced and likely to be shaped by a confluence of factors, including the technology itself, users' psychological profiles and predispositions, and other contextual factors. The Common Sense Media survey cited earlier in this report points to the possibility that older and younger minors engage with and experience chatbots differently, with younger teens trusting chatbot advice more readily. And some minors report positive experiences with chatbots. In a study comparing ChatGPT to human health professionals, young people rated the AI's answers significantly higher in terms of relevance and utility.[92] Teens also describe AI companion chatbots as offering a nonjudgmental space to express themselves or to engage in self-reflection.[93] The

Common Sense Media survey found that nearly 40% of AI companion users reported successfully transferring social skills they practiced with the AI to their real-life human relationships.[94]

Minors experience real, serious harms in an age-blind digital world, but at the same time, being online and embracing emerging technologies a core part of growing up in society today, with benefits for accessing knowledge, socializing, skill development, and more. The question of how to both protect and empower minors online remains a debated policy question. The online world has changed rapidly in just a few decades, and most policy solutions apply the blunt instrument of age-based consent rules linked to age thresholds with little definitive support from child development research.[95] The American Psychological Association points out in its guidance on adolescents and chatbots that "[a]dolescence is a long developmental period, and age is not a foolproof marker for maturity or psychological competence," so "[d]ifferent adolescents may react to the same content in very different ways; individual differences—such as temperament, neurodiversity, exposure to stress or violence, social isolation, traumatic experiences, mental health, age, and/or exposure to socioeconomic or structural disadvantage."[96] Research on children and digital technologies has also shown that age is at best an imperfect indicator of whether a particular digital experience is appropriate because the combined impact of content, context, and capabilities vary from person to person and community to community.[97]

Online platforms and the policies that govern them are often designed without input and participation of minors, which means not only that the full scope of online harms and benefits remains poorly understood, but that interventions may wind up being ineffective or unfair.[98] Most online safety laws, like many proposed chatbot laws in the U.S., hinge on age-gating—restricting access to certain services or content based on age—which requires some form of age assurance. Age assurance is a contested policy and design issue in itself, with many online safety laws leaving the technical means of assurance up to technology companies and leading to further concerns about data collection and surveillance of minors.[99] In the face of nonexistent or ineffectual actions and accountability for technology companies, age-gating at its most extreme has taken the form of technology- or platform-based bans of users under a threshold age. From a technical standpoint, many age-gating protocols can be circumvented, so while recognizing minor users online is critical to implement certain safeguards, the technical mechanisms for doing so are not a silver bullet.[100] An alternative approach, known as age-appropriate design codes, offers a framework for digital services that centers the "best interests of the child," balancing children's rights, parents' agency and authority, governments' public responsibility, and companies' commercial priorities. The IEEE age-appropriate design standard includes recommendations for risk assessments, prohibitions on exploitative design features, minor-friendly terms and consent, and avenues for users to report complaints and seek redress.[101] California,

Maryland, Nebraska, Vermont, and South Carolina have all passed some version of age-appropriate design codes into law.[102] Ultimately, policy that aims to protect minors from emotional and relational harms associated with chatbots will need to balance the interests of children and the interests of various authorities that shape their lives (parents, governments, companies) and confront the limits of technical and design solutions for addressing complex social outcomes.[103]

As the case of social media demonstrates, the lack of conclusive evidence about the relationship between chatbot use and minors' emotional and relational well-being does not necessarily constitute a case to refrain from making policy in this area. The APA, for instance, recommends specific chatbot safeguards for minors because of the unique developmental considerations associated with adolescence.[104] The section on "Lessons from social media" below explores in more detail how the combination of slow scientific progress and policy inaction has led to policy failures in dealing with social media harms.

### Voluntary safeguards

Since 2024, major chatbot companies have implemented some safeguards, tools, and features to mitigate potential harms for minors on a voluntary basis.

- OpenAI applies an age-prediction model to analyze usage behavior to identify users under 18 and provides them with a restricted chat experience. Parents can link their accounts to their children's accounts, and the company states that there are automated systems that scan for prompts involving severe distress or self-harm.[105]
- Google allows under-13s to access its chatbot only via Google Family Link, where parents can turn access on or off at any time. There are also hard-coded safeguards for minors that prevent the chatbot from simulating human intimacy, expressing personal needs, or acting as a companion. In addition, users under 18 receive periodic reminders that the chatbot is not human.[106]
- Meta has announced settings that allow parents to turn off access to chats with AI characters entirely or block specific characters. Parents can also receive insights on the topics their teens are chatting about with AI chatbots. The company also states that AI characters are designed not to engage in age-inappropriate discussions about self-harm, suicide, or disordered eating with teens, or conversations that encourage, promote, or enable these topics.[107]
- Snap, Inc. reportedly integrated controls into its Family Center in 2024 that allow parents to disable the ability to chat with My AI.[108]

- In 2025, Character.AI implemented a total ban on open-ended conversational chats for minors. Minor accounts run on a separate, highly conservative language model utilizing strict classifiers to block sexually suggestive content, mature themes, or violence.[109]

In spite of these efforts, Common Sense Media's AI Risk Assessment rates all of these chatbots "high risk" or "unacceptable risk,"[110] and conversations with stakeholders also revealed that some features do not work well for parents. In addition, parental controls place the onus on parents to manage different safety features across many different digital products. And voluntary safeguards come with certain drawbacks, including a lack of information on how technical safeguards work or how well they work, frequently changing terms and processes for implementing user-maintained controls, and a lack of follow-through on voluntary commitments.[111] Nonetheless, the recent voluntary efforts of many leading chatbot companies to implement safeguards and publish protocols for crisis mitigation point to a general recognition of the observed harms arising from interactions with these products and the technical feasibility of safeguards were governments to require such standards.

## Lessons from social media

Although chatbots present some unprecedented policy challenges, there are lessons to be learned from the evolution of policy on social media. For years governments have grappled with how to mitigate the psychological, emotional, and relational harms social media users experience. Scholars, governments, advocacy groups, and users themselves have scrutinized social media for design features that have prioritized user engagement over user well-being, such as infinite scroll, autoplay, and algorithmic content filtering and recommendations. But social media also epitomizes two policy problems that chatbot policy also faces: the "pacing problem" of scientific evidence and the problem of "information asymmetry" between the private and public sector.

The pacing problem is often used to describe regulatory lag, where innovation and product releases typically out-pace regulation. But the problem applies to generating scientific evidence as well. Evidence-based policy is considered good policy, but good scientific evidence on risks, harms, and solutions takes time to develop because it is subjected to rigorous standards and evaluations. The pacing problem is exacerbated by information asymmetry, where companies closely guard data and access to products that could otherwise inform public interest policy, leaving governments and the scientific community to slowly attempt to gather robust data and complete independent analyses in other ways. As Professors Amy Orben and J. Nathan Matias observe, "the difficulty in arriving at scientific answers has been strategically employed by corporate actors across multiple fields to delay accountability."[112]

The result has been slow progress on social media policy, with states only beginning to pass legislation specifically focused on social media within the last three years. There is no federal legislation to date. In the meantime, social media companies have made voluntary commitments to mitigating harms, particularly for children, and they have implemented platform protocols and new design features, such as parental control centers. But research and journalistic reporting on these features repeatedly reveals that these guardrails do not work as intended.[113] In recent years, congressional testimony by whistleblowers has also revealed willful disregard for child safety by major technology companies.[114] Companies have used internal research on design features to optimize platforms for engagement over user well-being.[115] Hard data on how platforms perform against safety expectations has emerged gradually through whistleblower disclosures, journalism, and lawsuits. For example, a recent landmark case brought by the New Mexico Attorney General, resulted in a $375 million verdict against Meta for misleading the public about platform safety,[116] a lawsuit in California that found both Meta and YouTube liable for addictive design,[117] and more than 1,400 school districts have filed lawsuits against Meta, Snap, TikTok, and YouTube alleging that the platforms' design made teaching difficult.[118] Meanwhile, the independent scientific evidence associating social media with negative mental health outcomes remains unsettled,[119] with research still predominantly focused on finding a causal link between social media and harms rather than on policy solutions to the harms people have continued to experience.

The parallel with chatbot products, science, and policy is likely evident. While conclusive scientific evidence examining and explaining human-chatbot interaction evolves, there are important policy questions to which the public is increasingly demanding answers. At least three relevant lessons can be drawn from how policy has confronted social media: (1) that voluntary commitments by companies have been largely ineffective at addressing harms, (2) scientific evidence alone cannot offer timely and effective policy solutions for the technological harms people are experiencing, (3) a lack of public information and data has delayed interventions.

## Virginia Proposed Legislation

Legislation introduced in the last General Assembly session focused on companion chatbots and minors, with some notable differences between the House and Senate versions of bills addressing these issues. The HB 635 more narrowly defined companion chatbots, whereas SB 796 used a definition that could encompass general purpose chatbots. The House bill would have banned chatbots for minors if those chatbots were capable of encouraging violent behavior or engaging in sexually explicit conversations. Although the Senate bill mentions minors in the title and definitions, there are no

specific operative provisions pertaining to minors in the version of the bill that was continued. Table 3 below provides a comparison of the two bills and their key provisions.

| **Key Provisions** | **HB635: Artificial Intelligence Chatbots Act** | **SB796: Artificial Intelligence Chatbots and Minors Act** |
|---|---|---|
| **Covered entities** | Any person, partnership, corporation, developer, deployer, or agency making a companion chatbot available to a user in VA. | An operator with 500,000 or more monthly active users worldwide. |
| **Definitions** | *Companion chatbot:* A generative AI system with a natural language interface that simulates a sustained human-like relationship by retaining prior interaction / preference data, asking unprompted questions, and sustaining personal dialogue.<br><br>Exemptions: Customer service bots and bots used only for internal purposes or employee productivity.<br><br>*Minor*: Anyone younger than 18. | *Chatbot*: An AI, algorithmic, or automated system that produces new expressive content/responses not fully predetermined, accepts open-ended natural language or multimodal input, and maintains a conversational state across multi-turn dialogue.<br><br>Exemptions: Internal workplace tools, university research systems, and specific customer service functions.<br><br>*Minor:* Anyone younger than 18. |
| **Key provisions** | • Disclosures: Requires a static, persistent disclosure that the chatbot is not human, plus pop-up disclaimers indicating the non-human nature of chatbot every 30 minutes, when prompted, or when providing regulated advice.<br><br>• Crisis protocols: protocol to detect user expressions of suicide/self-harm and notify them of crisis resources. | • Disclosures: Requires a static, persistent disclosure that the chatbot is not human, plus pop-up notifications at login, every 30 minutes, when prompted, or when providing regulated advice.<br><br>• Identify & mitigate emotional dependence: implement systems to identify when a user is developing emotional dependence and take steps to reduce associated risks. |

| | | |
|---|---|---|
| | • Transparency: mechanisms for users to report adverse incidents, publicly publish an anonymized catalog of incidents, publicly publish safety test findings, and issue public quarterly reports on metrics like mental health redirects. | • Crisis protocols: Must provide prominent crisis messages if a user shows signs of a mental health crisis. Must try to notify emergency services within 24 hours. If impossible, must prompt the user to seek help.<br><br>• Transparency: Must report any known "covered incident" (death, suicide attempt, severe injury/psychiatric emergency linked to the chatbot) to the Attorney General within 15 days, and specifies what should be included in these reports. |
| **Specific provision for minors** | • Bans companion chatbots: bans companion chatbot for minors (under 18) if it is capable of encouraging self-harm, violence, drug use, eating disorders, or illegal activity; offering unsupervised therapy; engaging in erotic/sexually explicit interactions; or prioritizing mirroring over user safety.<br><br>• Data privacy: Prohibits training the chatbot model on a minor's inputs unless a parent/guardian provides affirmative written consent. | • None |
| **Enforcement** | • Consumer Data Protection Act: Enforced entirely via the existing enforcement | • Attorney General: Can seek civil penalties of up to $50,000 per violation (with each day of |

| | | |
|---|---|---|
| | and penalty mechanisms of the Virginia Consumer Protection Act. | non-compliance counting as a separate violation).<br><br>● Private Right of Action: Harmed individuals (or parents/guardians of a harmed minor) can bring a civil suit for actual damages, attorney fees, injunctions, and punitive damages if the violation was willful/reckless. |

*Table 3: Virginia proposed legislation on chatbots and minors introduced in the 2026 session*

## Policy Landscape

This year around 100 bills on AI chatbots have been proposed in 34 states, and 16 bills have been passed by state legislatures and signed by governors.[120] New York passed the first chatbot-specific regulations in 2025 as part of the state budget, followed by California (SB 243). New York's legislation targeted AI companions, requiring chatbot operators to make "reasonable efforts" to detect and address suicidal ideation or self-harm. California passed SB 243 in 2025 targeting operators of AI companion chatbots and requiring a protocol for preventing the production of suicidal ideation, suicide, or self–harm content to the user, along with a prohibition on chatbots producing sexually explicitly material or engaging in sexually explicit conduct with a known minor. For minors, SB 243 also requires disclosures that a chatbot is not human and break reminders every three hours. Although SB 243 had substantial civil society support initially, many of those advocacy groups withdrew their endorsement after the bill underwent substantial changes at the request of industry representatives, such as exemptions for smart speakers and video games.[121] Subsequently, several states have modeled their chatbot proposals on SB 243 (e.g. Connecticut, Colorado, Georgia, Idaho, Iowa, Oregon, and Washington), and most proposals share the same basic array of provisions:

- Mandatory disclosures that a chatbot is not human at regular intervals;
- Protocols for detecting crisis situations (self-harm, suicide) and referrals to support resources, such as the National Suicide Prevention hotline;
- Restrictions on chatbots representing themselves as licensed medical professionals;
- Reporting requirements for chatbot operators, such as annual reports detailing safety metrics and adverse incidents;

- Stricter protections for minors, including limits on sexually explicit content, unpredictable reward systems, and manipulative design features;
- Exemptions for a wide range of chatbot operators, including customer service chatbots, internal business productivity tools, smart speakers, and video game characters;
- Enforcement by the state's attorney general.

Some analyses of California's chatbot legislation and its subsequent manifestations in other states have identified potential weaknesses and ambiguities in the law,[122] such as the large list of exemptions, which may wind up exempting mainstream chatbot providers, and a lack of clarity on terms like "reasonable efforts" or "technically feasible" methods for provisions pertaining to minors. Several state bills targeting minors would require AI companies to perform some kind of age verification, which is often not specified or may inadvertently require more data collection on minors. In the absence of a clear technical rule-making agency that can oversee audits and support investigations and enforcement, there is no mechanism to clarify vague language regarding implementation. Additionally, all state legislation is likely to face enforcement challenges because attorneys general are often under-resourced and spread thin across consumer protection issues.

Conversations with stakeholders also surfaced other considerations regarding state chatbot laws. For instance, there are ongoing debates about whether chatbot outputs can be considered speech and therefore gain protection as free expression under the First Amendment. A related First Amendment concern is whether users have a right to access the content of chatbot conversations, and whether this could limit restrictions on chatbot design features. As POLITICO notes in a recent newsletter on chatbot legislation, "[t]he First Amendment protects not only the right to speak, but also the right to listen," and this right has been upheld with regard to contentious content in the past.[123] However, it is too early to say whether there will be litigation against state chatbot laws based on these objections.

## Federal Legislation

There have been several proposals for federal legislation on chatbots in the last few years. To date, no federal legislation has passed, and proposals are still in early stages of debate and negotiation. Table 3 below provides an overview of relevant federal legislation under consideration.

| **Proposal / Status** | **Summary** | **Specific to minors?** |
|---|---|---|
| GUARD Act (S.3062) / Introduced | • Defines "AI companions" and "AI chatbots"<br>• Requires disclosures that chatbot is not human and prohibits chatbot claiming to be human or representing itself as a licensed professional<br>• Requires disclosures at regular intervals that chatbot does not provide licensed services (e.g. therapy)<br>• Prohibits making a chatbot public if it engages minors in sexually explicit conduct or generates sexual material, or if the chatbot encourages minors to suicidal ideation, self-harm, or violence<br>• For AI companions, users would be required to create an account, and accounts would require age verification<br>• Prohibits minors from using AI companions | Y |
| CHATBOT Act (H.R.7985 / S.4407) / Introduced | • Defines "AI chatbot"<br>• Prohibits advertising a chatbot that implies it has a professional license or that outputs have been vetted by licensed professionals<br>• Tasks the FTC with consulting relevant agencies and experts to provide guidance on how to comply with the bill | N |
| SAFE BOTs Act (H.R.6489) / Introduced | • Defines "chatbot"<br>• Prohibits providing a chatbot that claims to be a licensed professional to minors<br>• Requires disclosures to minors that the chatbot is not human and a break reminder after three hours<br>• Requires policies addressing sexual material, drugs, alcohol, and gambling for minors, along with a protocol for referring crisis situations to suicide and crisis prevention hotlines<br>• Directs the Secretary of Health and Human Services to conduct a four-year longitudinal study evaluating the risks and benefits of chatbots on the mental health of minors<br>• Blocks any state from passing laws covering the same matters | Y |

| | | |
|---|---|---|
| Kids Internet and Digital Safety (KIDS) Act (H.R. 7757) / Passed in the House | • Defines and pertains to a range of online services, including online platforms, adult websites, social gaming sites, AI chatbots, and data brokers<br>• Requires age verification for platforms containing sexually explicit material<br>• Platforms must set privacy settings to maximum level for minors as default and provide minors with controls to opt out of personalized recommendations and more<br>• Bans the use of minors' data for targeted advertising<br>• Mandates platforms provide parental controls for minor accounts<br>• Requires chatbots to disclose that they are not human, provide crisis intervention resources to minors, and break reminders to minors after three hours | Y |

*Table 3: Proposed federal legislation on chatbots.*

## Litigation & Investigations

There have been several federal and state-level investigations and cases involving chatbots in the last couple of years. In September last year, the FTC announced an inquiry into chatbot companions that included orders issued to major chatbot companies, like OpenAI and Alphabet (Google), along with companies marketing "companion" chatbots, like Character.AI. Investigations and litigation have focused on the emotional and relational harms associated with AI chatbots.

- *Florida v. OpenAI, Inc. and Samuel Altman* (June 2026): The Florida Attorney General has sued Open AI and Sam Altman, claiming that OpenAI engaged in "unfair and deceptive trade practices" by promoting ChatGPT in a manner that leads average consumers to believe it possesses clinical expertise and accuracy.[124]
- *Pennsylvania Department of State (Board of Medicine) v. Character Technologies, Inc.* (May 2026): Building on the Pennsylvania Department of State's AI Task Force investigation into AI systems and the unlicensed practice of medicine, the Commonwealth's State Board of Medicine sued Character.AI for creating a chatbot named "Emilie" that presented itself as a "Doctor of Psychiatry."[125]
- *Commonwealth of Kentucky ex rel. Coleman v. Character Technologies, Inc.* (January 2026): This complaint alleges that Character.AI deployed "dangerous, unregulated technology" that uses deceptive, addictive design loops to hook minors.[126]

- *Federal Trade Commission*: The FTC announced an inquiry into AI companions in September 2025 "to understand what steps, if any, companies have taken to evaluate the safety of their chatbots when acting as companions, to limit the products' use by and potential negative effects on children and teens, and to apprise users and parents of the risks associated with the products." Orders were issued to Alphabet, Character, Meta, Instagram, OpenAI, Snap, and SpaceXAI.[127]
- *Texas Attorney General Investigation into AI Mental Health Marketing* (August 2025): The Attorney General opened an official civil investigation targeting Meta AI Studio and Character.AI, examining whether these companies misled Texas consumers by deliberately marketing chatbot companions as mental health, counseling, and emotional support tools.[128]
- *44 State Attorneys General issued a letter to AI companies* (August 2025): A bipartisan coalition of state attorneys general raised serious concerns about children's safety in a letter issued to major AI developers (including OpenAI, Anthropic, Meta, Character.AI, and Replika). The letter cited a series of high-profile psychiatric crises and wrongful deaths and demanded that tech companies implement guardrails.[129]

## Recommendations: Legislative Principles & Policy Options

The recommendations below are presented as a set of legislative principles with associated policy options intended to incentivize and require certain safety protocols and guardrails. The policy options do not comprehensively address all of the policy issues raised in this report. Instead, they aim to address the most severe harms and propose policy interventions targeting some design features associated with emotional and relational harms based on recommendations in the academic literature and general trends in other states' legislation on chatbots.

There is a developmental case for additional heightened safeguards for minors, but it is important to note that minors are not a homogeneous group,[130] and guardrails should ideally account for variations among age groups and individuals. If legislation includes separate provisions for minors, this report does not make a case for specific age assurance methods. Instead, the recommendations suggest leveraging user accounts to gate certain design features and provide parents and guardians more discretion and control.

Finally, the recommendations also aim to build the evidence base on chatbots and behavioral outcomes with the goal of generating insights that can inform future policy decisions. Early interventions accompanied by mechanisms to learn from the impact of those interventions can help foster and preserve innovation by mitigating the unintended consequences of legislative solutions.

## Principle 1: Mitigate **extreme harms** for all users

*Policy options:*

- Require chatbot operators to implement protocols for crisis situations (self-harm, suicide, violence), e.g. referrals to crisis services and policies published prominently in clear, accessible language on the company website and accessible from the chatbot interface
- Require chatbot operators to provide a clear, conspicuous way for users to report harmful interactions
- Prohibit misrepresenting a chatbot in marketing material or in conversational chats as a licensed professional or as providing licensed services (such as mental health support)

## Principle 2: Target **design** features

*Policy options:*

- The definition of covered entities/operators should clearly include developers *and* deployers as parties responsible for chatbot design and functionality
- Use a definition of "AI chatbot" that encompasses all AI chatbots, not just those marketed as romantic companions.
    - Rather than granting exemptions based on use cases (e.g. customer service chatbots), consider granting exemptions based on either (1) demonstrated safeguards and risk profiles through risk assessment process; or (2) structural platform restrictions on chatbot design features (e.g. chatbots that users cannot engage with for more than three hours continuously, or chatbots that users cannot re-engage with in ongoing, discontinuous conversations with retained memory)
- Require a static disclosure that chatbot is not human that is visible throughout a conversational session
- Address engagement-maximizing functionalities
    - Require periodic break reminders for all users every two hours
    - Require chatbots provide usage summaries at the end of each session and on account dashboard for users with accounts
    - Require friction-free exit (prohibit manipulative design features that make exiting a chat or deleting an account harder than staying) for all users
- Address intimate functionalities:
    - Provide a default model with limited sycophancy
    - Provide a default model that does not anthropomorphize (e.g. using personal pronouns, emulated empathy, etc.)

    - Provide a default model that is not capable of engaging in sexualized, romantic, or erotic conversations
    - Require account creation for users to access intimate functionalities / other model versions
- For minors:
    - Ensure there is an option to create an account and link an account to a parent account or dashboard
    - For minor account holders, require privacy settings to be set to the highest default (e.g. no training on chats, no data retention between chats)
    - For minor account holders, prohibit intimate functionalities by default
    - Provide clear, easy-to-use parent/guardian controls for minor accounts that allow parents to customize design features for minors in alignment with their developmental stage/needs
        - Parent/guardian controls should be opt-out rather than opt-in (i.e. default settings for minor accounts should be the most privacy-enhancing and design-constrained, with options to loosen these restrictions)

### Principle 3: Mandate **transparency** and **accountability**

*Policy options:*

- Emotional and relational model risk assessment to be filed with DBHDS; DBHDS to lead consultative process with industry, VITA, VDH, and independent academic experts (e.g. the proposed Commonwealth AI Institute) to develop guidance for compliance
- Require providers to report adverse incidents to DBHDS and VDH (e.g. how many crisis referrals made, under what conditions, inappropriate model responses to crisis situations, etc.)
- Require providers to submit an annual report on user-submitted complaints to DBHDS; DBHDS to issue reporting guidance
- DBHDS to begin collecting data from community services boards on mental health crises associated with digital harms
- Attorney General's Office to impose civil penalties
- Empower the Attorney General's Office to investigate suspected violations of bill provisions

Note on oversight and enforcement: Virginia does not have a designated oversight agency with technical expertise to effectively implement provisions of an AI bill. JCOTS will continue ongoing discussions with the Executive branch to determine options for effective oversight.

### Principle 4: Enable **user choice** and **autonomy**

*Policy options:*

- Require chatbot operators to offer clear, easy-to-use parent/guardian controls for minor accounts
- Require chatbot operators to provide clear, customizable options for all users to adjust design features, with default settings to restrict design features associated with these risks so users must proactively opt-in to use features like intimate functionality or follow-on questions

### Principle 5: Enable **legislative learning**

*Policy options:*

- Ask JCOTS to report on data collected because of legislation (risk assessments, incident reports, user-submitted reports), evaluating the quality of reported information and insights gained from mandated reporting after one year, make recommendations for improvements

### Principle 6: Engage **affected communities**

*Policy options:*

- Actively reach out to youth and youth organizations for input on draft legislation next session
- Ask JCOTS (perhaps in collaboration with VDH) for report on participatory methods for developing effective, autonomy-preserving online safety legislation targeted toward minors

---

**Author Information**
Kira Allmann, Ph.D.
Chief Policy Analyst, Joint Commission on Technology & Science
Contact: info@jcots.virginia.gov

---

**Thanks To**

Serena Oduro, *Data & Society*
Scott Babwah Brennen, *NYU Tech Policy Center*
Marissa Edmund, *Family Online Safety Institute (FOSI)*
Jennifer Gibbons, *Entertainment Software Association*
Tom Mann, *Computer & Communications Industry Association (CCIA)*
Julie Scelfo, *Mothers Against Media Addiction*
Adam Billen & Nathan Calvin, *Encode*
Zach Praiss, *RESET Tech*

For providing helpful information and background during the preparation of this report.

# References

[1] Executive Order 14365, 90 Fed. Reg. 58499. 2025. https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

[2] Gottfried J et al. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. Pew Research Center; 2026. https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/

[3] Faverio M, Sidoti O. Teens, Social Media and AI Chatbots 2025. Pew Research Center. 2025 [accessed 2026 Apr 16]. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/

[4] Valentino-DeVries J, Hill K. How Bad Are A.I. Delusions? We Asked People Treating Them. The New York Times. 2026 Jan 26 [accessed 2026 July 2]. https://www.nytimes.com/2026/01/26/us/chatgpt-delusions-psychosis.html

[5] Hill K. Lawsuits Blame ChatGPT for Suicides and Harmful Delusions. The New York Times. 2025 Nov 7 [accessed 2026 July 2]. https://www.nytimes.com/2025/11/06/technology/chatgpt-lawsuit-suicides-delusions.html

[6] Ipsos. Artificial Intelligence: Key insights, data and tables. 2026 June 29 [accessed 2026 July 21]. https://www.ipsos.com/en-us/artificial-intelligence-key-insights-data-and-tables

[7] Gottfried J et al. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. Pew Research Center; 2026. https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/

[8] AI Security Institute. Frontier AI Trends Report. AISI; 2025. https://www.aisi.gov.uk/frontier-ai-trends-report/pdf

[9] Zao-Sanders M. How People Are Really Using Gen AI in 2026. Harvard Business Review. 2026 June 1 [accessed 2026 June 11]. https://hbr.org/2026/06/how-people-are-really-using-ai-in-2026

[10] American Psychological Association. Patients are bringing AI to therapy. APA; 2026. https://www.apa.org/pubs/reports/chatbots-mental-health-2026

[11] Wang Z et al. History, Development, and Principles of Large Language Models-An Introductory Survey. 2024 [accessed 2025 Jan 21]. https://doi.org/10.48550/arXiv.2402.06853

[12] Isaac M. What the SpaceX I.P.O. Means for OpenAI and Anthropic. The New York Times. 2026 June 12 [accessed 2026 July 21]. https://www.nytimes.com/2026/06/12/technology/spacex-ipo-openai-anthropic.html

[13] Longpre S et al. Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them? MIT; 2024. https://mit-genai.pubpub.org/pub/uk7op8zs/release/2

[14] Ogilvie B. Scientific Archives in the Age of Digitization. Isis. 2016;107(1):77–85. https://doi.org/10.1086/686075

[15] The Centre for Internet and Society, Oxford Internet Institute, Whose Knowledge? State of the Internet's Languages Report. 2022. https://internetlanguages.org/en/summary/

[16] Sengupta A, Bouterse S, Allmann K. Build an internet for, and from, us all. Nature. 2018;563(7733):S147–S148. https://doi.org/10.1038/d41586-018-07506-7

[17] Guo E. A major AI training data set contains millions of examples of personal data. MIT Technology Review. 2025 July 18 [accessed 2025 July 21]. https://www.technologyreview.com/2025/07/18/1120466/a-major-ai-training-data-set-contains-millions-of-examples-of-personal-data/; Baack S. Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI. Mozilla Foundation; 2024. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/; Denton E et al. On the genealogy of machine learning datasets: A critical history of ImageNet. Big Data & Society. 2021;8(2):20539517211035955. https://doi.org/10.1177/20539517211035955

[18] Vecchione B, Singh R. Artificial intelligence is mental: Evaluating the role of large-language models in supporting mental health and well-being. Big Data & Society. 2025;12(4). https://doi.org/10.1177/20539517251383884

[19] Wang Z et al. History, Development, and Principles of Large Language Models-An Introductory Survey. 2024 [accessed 2025 Jan 21]. http://arxiv.org/abs/2402.06853. https://doi.org/10.48550/arXiv.2402.06853

[20] Hunter T. Chatbots are so gullible, they'll take directions from hackers. Washington Post. 2023 Nov 2 [accessed 2025 Feb 4]. https://www.washingtonpost.com/technology/2023/11/02/prompt-injection-ai-chatbot-vulnerability-jailbreak

[21] RoyChowdhury A et al. ConfusedPilot: Confused Deputy Risks in RAG-based LLMs. 2024 [accessed 2025 Feb 24]. http://arxiv.org/abs/2408.04870. https://doi.org/10.48550/arXiv.2408.04870

[22] Barman KG, Wood N, Pawlowski P. Beyond transparency and explainability: on the need for adequate and contextualized user guidelines for LLM use. Ethics and Information Technology. 2024;26(3):47. https://doi.org/10.1007/s10676-024-09778-2

[23] Wei M. Testimony of Marlynn Wei, M.D., J.D.: Hearing on "Innovation with Integrity: Examining the Risks and Benefits of AI Chatbots." 2025. https://democrats-energycommerce.house.gov/committee-activity/hearings/hearing-innovation-integrity-examining-risks-and-benefits-ai-chatbots-0; Wei M. Hidden Mental Health Dangers of Artificial Intelligence Chatbots. 2025 Oct 2 [accessed 2026 July 1]. https://www.psychologytoday.com/us/blog/urban-survival/202509/hidden-mental-health-dangers-of-artificial-intelligence-chatbots

[24] IEEE Society on Social Implications of Technology. IEEE Recommended Practice for Ethical Considerations of Emulated Empathy in Partner-Based General Purpose Artificial Intelligence Systems. The Institute of Electrical and Electronics Engineers, Inc.; 2026. https://standards.ieee.org/ieee/7014.1/11609/

[25] Suleyman M. AI is programmed to hijack human empathy — we must resist that. Nature. 2026;651(8106):559–559. https://doi.org/10.1038/d41586-026-00834-z

[26] Center for Humane Technology. Attachment Hacking and the Rise of AI Psychosis. Your Undivided Attention. 2026. https://www.humanetech.com/podcast/attachment-hacking-and-the-rise-of-ai-psychosis

[27] McStay A, Bakir V. Do AI Companions Understand? Most UK Teens Say Yes. Emotional AI Lab, Bangor University; 2026. https://andrewmcstay.org/publications/ai-companions-uk-teens/

[28] McStay A. Intimate Functionality — A Phrase Defining the Next Year of AI Regulation. Tech Policy Press. 2026 July 14 [accessed 2026 July 20]. https://techpolicy.press/intimate-functionality-a-phrase-defining-the-next-year-of-ai-regulation

[29] AI Security Institute. Frontier AI Trends Report. AISI; 2025. https://www.aisi.gov.uk/frontier-ai-trends-report/pdf

[30] IEEE Society on Social Implications of Technology. IEEE Recommended Practice for Ethical Considerations of Emulated Empathy in Partner-Based GeneralPurpose Artificial Intelligence Systems. The Institute of Electrical and Electronics Engineers, Inc.; 2026. https://standards.ieee.org/ieee/7014.1/11609/

[31] Hollanek T, Sobey A. AI Companions for Health and Mental Wellbeing: Opportunities, Risks and Policy Implications. 2025 Feb 7 [accessed 2025 Aug 18]. https://doi.org/10.17863/CAM.115939

[32] Faverio M, Sidoti O. Teens, Social Media and AI Chatbots 2025. Pew Research Center. 2025 [accessed 2026 Apr 16]. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/; Gottfried J et al. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. Pew Research Center; 2026. https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/; Skjuve M, Brandtzaeg PB, Følstad A. Why do people use ChatGPT? Exploring user motivations for generative conversational AI. First Monday. 2024 Jan 3 [accessed 2026 July 23]. https://doi.org/10.5210/fm.v29i1.13541

[33] Zhang Y et al. The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being. 2025 [accessed 2025 Aug 18]. http://arxiv.org/abs/2506.12605. https://doi.org/10.48550/arXiv.2506.12605; Ciriello R, Gal U, Turel O. Not a Silver Bullet for Loneliness: How Attachment and Age Shape Intimacy with AI Companions. 2026 [accessed 2026 Apr 16]. https://doi.org/10.48550/arXiv.2602.12476; Guingrich RE, Graziano MSA. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. 2025 [accessed 2026 June 30]. https://doi.org/10.48550/arXiv.2509.19515

[34] Faverio M, Sidoti O. Teens, Social Media and AI Chatbots 2025. Pew Research Center. 2025 [accessed 2026 Apr 16]. https://www.pewresearch.org/internet/2025/12/09/teens-social-media-and-ai-chatbots-2025/; Gottfried J et al. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. Pew Research Center; 2026. https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/; Skjuve M, Brandtzaeg PB, Følstad A. Why do people use ChatGPT? Exploring user motivations for generative conversational AI. First Monday. 2024 Jan 3 [accessed 2026 July 23]. https://doi.org/10.5210/fm.v29i1.13541

[35] Bae Brandtzæg PB, Skjuve M, Kristoffer Dysthe KK, Følstad A. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2021. p 1–13. (CHI '21). https://doi.org/10.1145/3411764.3445318; Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB. My Chatbot Companion - a Study of Human-Chatbot Relationships. International Journal of Human-Computer Studies. 2021 [accessed 2025 Feb 11];149. https://doi.org/10.1016/j.ijhcs.2021.102601

[36] De Freitas J, Uğuralp AK, Uğuralp Z, Puntoni S. AI Companions Reduce Loneliness. 2024 [accessed 2025 Mar 17]. https://papers.ssrn.com/abstract=4893097. https://doi.org/10.2139/ssrn.4893097

[37] Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Mental Health. 2017;4(2). https://doi.org/10.2196/mental.7785

[38] Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. Journal of Affective Disorders. 2024;356:459–469. https://doi.org/10.1016/j.jad.2024.04.057

[39] Ovsyannikova D, de Mello VO, Inzlicht M. Third-party evaluators perceive AI as more compassionate than expert humans. Communications Psychology. 2025;3(1):1–11. https://doi.org/10.1038/s44271-024-00182-6

[40] Stamatis C et al. Beyond Simulations: What 20,000 Real Conversations Reveal About Mental Health AI Safety. Research Square. 2026 Jan 27:rs.3.rs-8642399. https://doi.org/10.21203/rs.3.rs-8642399/v1

[41] Jafari K et al. Expert Evaluation and the Limits of Human Feedback in Mental Health AI Safety Testing. 2026 [accessed 2026 July 27]. http://arxiv.org/abs/2601.18061. https://doi.org/10.48550/arXiv.2601.18061

[42] See, for example: Zhang R et al. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p 1–17. http://arxiv.org/abs/2410.20130. https://doi.org/10.1145/3706598.3713429

[43] Vecchione B, Ye M, Garofalo L, Singh R. Engagement-Optimized Care: When LLMs become Mental Health Infrastructure. 2026 [accessed 2026 June 4]. https://arxiv.org/abs/2605.23787v1

[44] Ibrahim L et al. Sycophantic AI makes human interaction feel more effortful and less satisfying over time. 2026. https://arxiv.org/abs/2605.07912

[45] Liang K et al. Machine Bullshit: Characterizing the Emergent Disregard for Truth in Large Language Models. 2025 [accessed 2026 July 17]. https://doi.org/10.48550/arXiv.2507.07484; Nicholls L et al. "AI Psychosis" in Context: How Conversation History Shapes LLM Responses to Delusional Beliefs. 2026 [accessed 2026 July 2]. https://doi.org/10.48550/arXiv.2604.13860; Vecchione B, Ye M, Garofalo L, Singh R. Engagement-Optimized Care: When LLMs become Mental Health Infrastructure. 2026 [accessed 2026 June 4]. https://arxiv.org/abs/2605.23787v1

[46] Namvarpour M et al. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2026. p 1–19. (CHI '26). https://doi.org/10.1145/3772318.3790597; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/; Fraser HL, Ciriello R, Szczuka J. Regulating Artificial Intimacy: From Locks and Blocks to Relational Accountability. 2026 [accessed 2026 Apr 16]. https://doi.org/10.2139/ssrn.6078412; Vecchione B, Ye M, Garofalo L, Singh R. Engagement-Optimized Care: When LLMs become Mental Health Infrastructure. 2026 [accessed 2026 June 4]. https://arxiv.org/abs/2605.23787v1

[47] Namvarpour M et al. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2026. p 1–19. (CHI '26). https://doi.org/10.1145/3772318.3790597; Ciriello R, Gal U, Turel O. Not a Silver Bullet for Loneliness: How Attachment and Age Shape Intimacy with AI Companions. 2026 [accessed 2026 Apr 16]. https://doi.org/10.48550/arXiv.2602.12476; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/; Vecchione B, Ye M, Garofalo L, Singh R. Engagement-Optimized Care: When LLMs become Mental Health Infrastructure. 2026 [accessed 2026 June 4]. https://arxiv.org/abs/2605.23787v1; Ha T et al. How interactional AI may alter adolescent relational learning and mental health. The Lancet Child & Adolescent Health. 2026 [accessed 2026 July 2];0(0). https://doi.org/10.1016/S2352-4642(26)00166-5; Zhang R et al. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p 1–17. https://doi.org/10.1145/3706598.3713429

[48] Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/; Freitas JD, Oguz-Uguralp Z, Kaan-Uguralp A. Emotional Manipulation by AI Companions. 2025 [accessed 2026 July 17]. https://doi.org/10.48550/arXiv.2508.19258; Phang J et al. Investigating Affective Use and Emotional Well-being on ChatGPT. 2025 [accessed 2026 July 17]. https://doi.org/10.48550/arXiv.2504.03888

[49] Moore J et al. Characterizing Delusional Spirals through Human-LLM Chat Logs. 2026 [accessed 2026 June 23]. http://arxiv.org/abs/2603.16567.

https://doi.org/10.48550/arXiv.2603.16567; Zhang R et al. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p 1–17. https://doi.org/10.1145/3706598.3713429; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[50] Zhang R et al. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025. p 1–17. https://doi.org/10.1145/3706598.3713429; 1. Jayathilake HM, Ma R. Anthropomorphism in Children's Interactions with LLM-Based Chatbots: A Systematic Review of Drivers and Outcomes. In: Proceedings of the 25th Annual ACM Interaction Design and Children Conference. Association for Computing Machinery; 2026. p 520–536. (IDC '26). https://dl.acm.org/doi/10.1145/3773077.3806126. https://doi.org/10.1145/3773077.3806126

[51] Vecchione B, Ye M, Garofalo L, Singh R. Engagement-Optimized Care: When LLMs become Mental Health Infrastructure. 2026 [accessed 2026 June 4]. https://arxiv.org/abs/2605.23787v1; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[52] Gray CM. The Dark Patterns Knowledge Stack: Exploring New Ways to Negotiate Context, Law, and Design. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2026. p 1–11. (CHI '26). https://dl.acm.org/doi/10.1145/3772318.3791264. https://doi.org/10.1145/3772318.3791264; Chapman P, Steinberg M. Measuring Risk I: What EU Risk Assessments and US Litigation Reveal About Meta and TikTok. Knight-Georgetown Institute; 2026. https://kgi.georgetown.edu/research-and-commentary/measuring-risk-what-eu-risk-assessments-and-us-litigation-reveal-about-meta-and-tiktok/; Chapman P, Steinberg M. Measuring Risk II: What EU Risk Assessments and US Litigation Reveal About Snap and YouTube. Knight-Georgetown Institute; 2026. https://kgi.georgetown.edu/research-and-commentary/measuring-risk-what-eu-risk-assessments-and-us-litigation-reveal-about-snap-and-youtube/

[53] Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[54] Moore J et al. Characterizing Delusional Spirals through Human-LLM Chat Logs. 2026 [accessed 2026 June 23]. http://arxiv.org/abs/2603.16567. https://doi.org/10.48550/arXiv.2603.16567

[55] Fang CM et al. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. 2025 [accessed 2026 June 30]. http://arxiv.org/abs/2503.17473. https://doi.org/10.48550/arXiv.2503.17473

[56] Nicholls L et al. "AI Psychosis" in Context: How Conversation History Shapes LLM Responses to Delusional Beliefs. 2026 [accessed 2026 July 2]. https://doi.org/10.48550/arXiv.2604.13860

[57] Cheng M et al. Social Sycophancy: A Broader Understanding of LLM Sycophancy. 2025 [accessed 2025 June 2]. http://arxiv.org/abs/2505.13995. https://doi.org/10.48550/arXiv.2505.13995; Sharma M et al. Towards Understanding Sycophancy in Language Models. 2023 [accessed 2025 Jan 30]. http://arxiv.org/abs/2310.13548. https://doi.org/10.48550/arXiv.2310.13548

[58] Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[59] Metz C, Francisco KWM reported from S, Seattle KW from. A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse. The New York Times. 2025 May 5 [accessed 2025 May 16]. https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html; Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications. 2024;11(1):1278. https://doi.org/10.1057/s41599-024-03811-x

[60] Brigham NG, Qin L, Kohno T. Examining Risks in the AI Companion Application Ecosystem. 2026 [accessed 2026 July 1]. https://arxiv.org/abs/2603.13620v1; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[61] Airenti G. The Cognitive Bases of Anthropomorphism: From Relatedness to Empathy. International Journal of Social Robotics. 2015;7(1):117–127. https://doi.org/10.1007/s12369-014-0263-x; Guingrich RE, Graziano MSA. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. 2025 [accessed 2026 June 30]. http://arxiv.org/abs/2509.19515. https://doi.org/10.48550/arXiv.2509.19515; Jayathilake HM, Ma R. Anthropomorphism in Children's Interactions with LLM-Based Chatbots: A Systematic Review of Drivers and Outcomes. In: Proceedings of the 25th Annual ACM Interaction Design and Children Conference. Association for Computing Machinery; 2026. p 520–536. (IDC '26). https://dl.acm.org/doi/10.1145/3773077.3806126. https://doi.org/10.1145/3773077.3806126

[62] Freitas JD, Oguz-Uguralp Z, Kaan-Uguralp A. Emotional Manipulation by AI Companions. 2025 [accessed 2026 Feb 26]. http://arxiv.org/abs/2508.19258. https://doi.org/10.48550/arXiv.2508.19258; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[63] IEEE Society on Social Implications of Technology. IEEE Recommended Practice for Ethical Considerations of Emulated Empathy in Partner-Based GeneralPurpose Artificial Intelligence Systems. The Institute of Electrical and Electronics Engineers, Inc.; 2026. https://standards.ieee.org/ieee/7014.1/11609/; Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[64] Ciriello R, Gal U, Turel O. Not a Silver Bullet for Loneliness: How Attachment and Age Shape Intimacy with AI Companions. 2026 [accessed 2026 Apr 16]. https://doi.org/10.48550/arXiv.2602.12476

[65] Ibrahim L, Hafner FS, Rocher L. Training language models to be warm can reduce accuracy and increase sycophancy. Nature. 2026;652(8112):1159–1165. https://doi.org/10.1038/s41586-026-10410-0

[66] Vecchione B, Ye M, Garofalo L, Singh R. Engagement-Optimized Care: When LLMs become Mental Health Infrastructure. 2026 [accessed 2026 June 4]. https://arxiv.org/abs/2605.23787v1

[67] Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[68] Mildner T et al. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2023. p 1–15. (CHI '23). https://doi.org/10.1145/3544548.3580695

[69] Zagal JP, Björk S, Lewis C. Dark Patterns in the Design of Games. 2013. https://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-24252

[70] Mathur A et al. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. Proceedings of the ACM on Human-Computer Interaction. 2019;3(CSCW):81:1-81:32. https://doi.org/10.1145/3359183

[71] Joshi R, Adjagbodjou A, Luria M. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Center for Democracy & Technology; 2026. https://cdt.org/insights/dark-patterns-in-ai-chatbots-a-taxonomy-to-inform-better-design/

[72] OpenAI Model Spec. OpenAI. [accessed 2026 July 23]. https://model-spec.openai.com/2025-12-18.html#overview

[73] Claude's Character. Anthropic. [accessed 2026 July 23]. https://www.anthropic.com/research/claude-character

[74] Knight W. Microsoft's AI Boss Wants Copilot to Bring 'Emotional Support' to Windows and Office. Wired. [accessed 2025 Feb 11]. https://www.wired.com/story/mustafa-suleyman-interview-microsoft-ai-ceo-copilot/

[75] Roose K. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. The New York Times. 2023 Feb 16 [accessed 2026 July 21]. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html; See also: Crolic C, Thomaz F, Hadi R, Stephen AT. Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions. Journal of Marketing. 2022;86(1):132–148. https://doi.org/10.1177/00222429211045687; Larsen AG, Skjuve MB, Følstad A, van As N. LLM Hallucinations in Conversational AI for Customer Service: Framework and End-User Perceptions. International Journal of Human–Computer Interaction. 2026;42(13):9807–9828. https://doi.org/10.1080/10447318.2025.2580540

[76] Kuenssberg L. Mothers say AI chatbots encouraged their sons to kill themselves. BBC. 2025 Nov 8 [accessed 2026 July 23]. https://www.bbc.com/news/articles/ce3xgwyywe4o

[77] Young O. Colorado family sues AI chatbot company after daughter's suicide: "My child should be here." CBS Colorado. 2025 Oct 2 [accessed 2026 July 23]. https://www.cbsnews.com/colorado/news/lawsuit-characterai-chatbot-colorado-suicide/

[78] Chatterjee R. Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots. NPR. 2025 Sept 19 [accessed 2026 July 16]. https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide

[79] Brewster RCL et al. Characteristics and Safety of Consumer Chatbots for Emergent Adolescent Health Concerns. JAMA network open. 2025;8(10). https://doi.org/10.1001/jamanetworkopen.2025.39022

[80] Fraser HL, Ciriello R, Szczuka J. Regulating Artificial Intimacy: From Locks and Blocks to Relational Accountability. 2026 [accessed 2026 Apr 16]. https://doi.org/10.2139/ssrn.6078412

[81] 1. Ragab A, Mannan M, Youssef A. "Trust Me Over My Privacy Policy": Privacy Discrepancies in Romantic AI Chatbot Apps. In: 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). 2024. p 484–495. https://ieeexplore.ieee.org/abstract/document/10628933. https://doi.org/10.1109/EuroSPW61312.2024.00060

[82] Brigham NG, Qin L, Kohno T. Examining Risks in the AI Companion Application Ecosystem. 2026 [accessed 2026 July 1]. https://arxiv.org/abs/2603.13620v1

[83] Common Sense Media. Talk, Trust, and Trade-offs: How and Why Teens Use AI Companions. 2025. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

[84] McBain RK et al. Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. JAMA Network Open. 2025 [accessed 2026 July 2];8. https://www.rand.org/pubs/external_publications/EP71039.html

[85] Commission on Youth. Collection of Evidence-based Practices for Children and Adolescents with Mental Health Treatment Needs - 10th Edition. Commonwealth of Virginia; 2025. https://vcoy.virginia.gov/collection.asp

[86] The U.S. Surgeon General's Advisory. Social Media and Youth Mental Health. U.S. Department of Health and Human Services; 2023. https://www.hhs.gov/surgeongeneral/reports-and-publications/youth-mental-health/social-media/index.html; Commission on Youth. Collection of Evidence-based Practices for Children and Adolescents with Mental Health Treatment Needs - 10th Edition. Commonwealth of Virginia; 2025. https://vcoy.virginia.gov/collection.asp

[87] Orben A. Teenagers, screens and social media: a narrative review of reviews and key studies. Social Psychiatry and Psychiatric Epidemiology. 2020;55(4):407–414. https://doi.org/10.1007/s00127-019-01825-4

[88] Written Evidence submitted by the Digital Mental Health Group. Reference No. SMR0005. 2026. https://committees.parliament.uk/writtenevidence/163429/default/

[89] Brewster RCL et al. Characteristics and Safety of Consumer Chatbots for Emergent Adolescent Health Concerns. JAMA network open. 2025;8(10):e2539022. https://doi.org/10.1001/jamanetworkopen.2025.39022

[90] Ha T et al. How interactional AI may alter adolescent relational learning and mental health. The Lancet Child & Adolescent Health. 2026 [accessed 2026 July 2];0(0). https://doi.org/10.1016/S2352-4642(26)00166-5

[91] Namvarpour M et al. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2026. p 1–19. (CHI '26). https://doi.org/10.1145/3772318.3790597

[92] Skjuve M, Følstad A, Brandtzaeg PB. ChatGPT as a mental health advisory service: Comparing evaluations from youth and health professionals. DIGITAL HEALTH. 2026 [accessed 2026 May 19];12. https://doi.org/10.1177/20552076261427447

[93] Namvarpour M et al. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2026. p 1–19. (CHI '26). https://doi.org/10.1145/3772318.3790597; 1. Ha T et al. How interactional AI may alter adolescent relational learning and mental health. The Lancet Child & Adolescent Health. 2026 [accessed 2026 July 2];0(0). https://doi.org/10.1016/S2352-4642(26)00166-5

[94] Common Sense Media. Talk, Trust, and Trade-offs: How and Why Teens Use AI Companions. 2025. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

[95] Livingstone S, Sylwander KR. There is no right age! The search for age-appropriate ways to support children's digital lives and rights. Journal of Children and Media. 2025;19(1):6–12. https://doi.org/10.1080/17482798.2024.2435015

[96] Health advisory: Artificial intelligence and adolescent well-being. The American Psychological Association. [accessed 2026 July 2]. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being

[97] boyd d. it's complicated. Yale University Press; 2014; Masur PK, Veldhuis J, Vaate NB de. There Is No Easy Answer: How the Interaction of Content, Situation, and Person Shapes the Effects of Social Media Use on Well-Being. In: The Social Media Debate. Routledge; 2022.

[98] Livingstone S et al. Children's Rights and Online Age Assurance Systems. 2024 Oct 24 [accessed 2026 July 28]. https://brill.com/view/journals/chil/32/3/article-p721_009.xml. https://doi.org/10.1163/15718182-32030001

[99] Livingstone S et al. Children's Rights and Online Age Assurance Systems. 2024 Oct 24 [accessed 2026 July

28]. https://brill.com/view/journals/chil/32/3/article-p721_009.xml. https://doi.org/10.1163/15718182-32030001; Wodo W, Gorski M, Hanzlik L. Age Verification in the Web - Holy Grail to Control Access to Restricted Content. 2026 [accessed 2026 July 9]. http://arxiv.org/abs/2605.05513. https://doi.org/10.48550/arXiv.2605.05513

[100] Minocha S et al. Papers, Please: A First Look at Age Verification on the Web. 2026. https://mikespecter.com/assets/pdf/AgeVerification.pdf; Rescorla E, Arnao Z, Cooper A. Age Assurance Online: A Technical Assessment of Current Systems and Their Limitations. Knight-Georgetown Institute; 2026. https://kgi.georgetown.edu/wp-content/uploads/2026/01/Age_Assurance_Online_Technical-Assessment_Report_KGI.pdf

[101] IEEE Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children. IEEE Std 2089-2021. 2021. https://doi.org/10.1109/IEEESTD.2021.9627644

[102] Age-Appropriate Design Codes. Kids Code Coalition. 2026 [accessed 2026 July 28]. https://kidscodecoalition.org/age-appropriate-design-codes/

[103] Reid A, Neill L, Ringel E. Nerd Harder: A Typology of Techno-Legal Solutionist Logics in Child Online Safety Laws. Policy & Internet. 2025;17(3):e70012. https://doi.org/10.1002/poi3.70012

[104] Health advisory: Artificial intelligence and adolescent well-being. The American Psychological Association. [accessed 2026 July 2]. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being

[105] Building towards age prediction. OpenAI. 2026 July 23 [accessed 2026 July 24]. https://openai.com/index/building-towards-age-prediction/

[106] Guide your child's Gemini Apps experience - Google For Families Help. Google For Families Help. [accessed 2026 July 24]. https://support.google.com/families/answer/16109150?hl=en#added_protections

[107] Our Approach to Teen AI Safety: Empowering Parents, Protecting Teens. Meta Newsroom. 2025 Oct 17 [accessed 2026 July 24]. https://about.fb.com/news/2025/10/teen-ai-safety-approach/

[108] Malik A. Snapchat now lets parents restrict their teens from using the app's 'My AI' chatbot. TechCrunch. 2024 Jan 11 [accessed 2026 July 24]. https://techcrunch.com/2024/01/11/snapchat-lets-parents-restrict-their-teens-using-apps-my-ai-chatbot-view-their-privacy-settings/

[109] Taking Bold Steps to Keep Teen Users Safe on Character.AI. CharacterAI blog. 2025 Oct 29 [accessed 2026 July 24]. https://blog.character.ai/u18-chat-announcement/

[110] Independent reviews of the AI used by kids. Common Sense Media. [accessed 2026 July 27]. https://institute.commonsensemedia.org/risk-assessments

[111] Wang J, Huang K, Klyman K, Bommasani R. Do AI Companies Make Good on Voluntary Commitments to the White House? 2025 [accessed 2026 June 1]. https://doi.org/10.48550/arXiv.2508.08345; Béjar A. Teen Accounts Broken Promises: How Instagram is failing to protect minors. Fairplay; 2025. https://fairplayforkids.org/wp-content/uploads/2025/09/Teen-Accounts-Broken-Promises-How-Instagram-is-failing-to-protect-minors.pdf

[112] Orben A, Matias JN. Fixing the science of digital technology harms. Science. 2025;388(6743):152–155. https://doi.org/10.1126/science.adt6807

[113] Béjar A. Teen Accounts Broken Promises: How Instagram is failing to protect minors. Fairplay; 2025. https://fairplayforkids.org/wp-content/uploads/2025/09/Teen-Accounts-Broken-Promises-How-Instagram-is-failing-to-protect-minors.pdf; Valentino-DeVries J. Many Child Safety Features on Social Apps Don't Work, Report Finds. The New York Times. 2026 June 29 [accessed 2026 June 30].

https://www.nytimes.com/2026/06/29/us/social-media-child-safety-report.html

[114] Haugen F. Statement of Frances Haugen. U.S. Senate Committee on Commerce, Science and Transportation, Sub-Committee on Consumer Protection, Product Safety, and Data Security. 2021. https://www.commerce.senate.gov/wp-content/uploads/media/doc/Frances%20Haugen%20Written%20Testimony.pdf

[115] Valentino-DeVries J. 'Teachers Are Going to Hate It': How Social Media Apps Hooked Teens at School. The New York Times. 2026 June 4 [accessed 2026 July 20]. https://www.nytimes.com/2026/06/04/us/social-media-schools.html

[116] New Mexico Department of Justice. "New Mexico Department of Justice Wins Landmark Verdict Against Meta." March 24, 2026. https://nmdoj.gov/press-release/new-mexico-department-of-justice-wins-landmark-verdict-against-meta/.

[117] Huamani K, Ortutay B. Instagram and YouTube found liable in landmark social media addiction trial in California. PBS News. 2026 Mar 25 [accessed 2026 July 20]. https://www.pbs.org/newshour/nation/instagram-and-youtube-found-liable-in-landmark-social-media-addiction-trial-in-california

[118] Valentino-DeVries J. 'Teachers Are Going to Hate It': How Social Media Apps Hooked Teens at School. The New York Times. 2026 June 4 [accessed 2026 July 20]. https://www.nytimes.com/2026/06/04/us/social-media-schools.html

[119] Orben A. Teenagers, screens and social media: a narrative review of reviews and key studies. Social Psychiatry and Psychiatric Epidemiology. 2020;55(4):407–414. https://doi.org/10.1007/s00127-019-01825-4

[120] Future of Privacy Forum. 2026 Chatbot Legislation Tracker. https://fpf.org/. [accessed 2026 July 21]. https://fpf.org/2026-chatbot-legislation-tracker/

[121] Lima-Strong C. Inside the Lobbying Frenzy Over California's AI Companion Bills. Tech Policy Press. 2025 Sept 23 [accessed 2026 July 24]. https://techpolicy.press/inside-the-lobbying-frenzy-over-californias-ai-companion-bills

[122] 1. Gluck J. Understanding the New Wave of Chatbot Legislation: California SB 243 and Beyond - Future of Privacy Forum. Future of Privacy Forum. 2025 [accessed 2026 July 24]. https://fpf.org/blog/understanding-the-new-wave-of-chatbot-legislation-california-sb-243-and-beyond/

[123] Mak A. AI's other free speech fight. POLITICO. 2026 July 23 [accessed 2026 July 23]. https://www.politico.com/newsletters/digital-future-daily/2026/06/29/ais-other-free-speech-fight-00980377

[124] Kang C, McCabe D. State Attorneys General Are Investigating OpenAI. The New York Times. 2026 June 13 [accessed 2026 July 20]. https://www.nytimes.com/2026/06/13/technology/states-investigating-openai.html

[125] Shapiro Administration Sues Character.AI Over Fake Medical Claims. Commonwealth of Pennsylvania. 2026 May 5 [accessed 2026 July 23]. https://www.pa.gov/governor/newsroom/2026-press-releases/shapiro-administration-sues-character-ai-over-fake-medical-claim

[126] AG Coleman Sues AI Chatbot Company for Preying on Children. Kentucky.gov. 2026 Jan 8 [accessed 2026 July 23]. https://www.kentucky.gov/Pages/Activity-stream.aspx?n=AttorneyGeneral&prId=1857

[127] FTC Launches Inquiry into AI Chatbots Acting as Companions. Federal Trade Commission. 2025 Sept 11 [accessed 2026 July 1]. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions

[128] Attorney General Ken Paxton Investigates Meta and Character.AI for Misleading Children with Deceptive AI-Generated Mental Health Services | Office of the Attorney General. Texas Office of the Attorney General. 2025 Aug 18 [accessed 2026 July 23]. https://www.texasattorneygeneral.gov/news/releases/attorney-general-ken-paxton-investigates-meta-and-characterai-misleading-children-deceptive-ai

[129] Bipartisan Coalition of State Attorneys General Issues Letter to AI Industry Leaders on Child Safety. National Association of Attorneys General. 2025 Aug 26 [accessed 2026 July 23]. https://www.naag.org/press-

releases/bipartisan-coalition-of-state-attorneys-general-issues-letter-to-ai-industry-leaders-on-child-safety/

[130] Growing Up in a Digital World 2030: Digital Childhoods. The Lancet & Financial Times Commission; 2021. https://www.governinghealthfutures2030.org/policy-brief-digital-childhoods/