Date of Hearing: April 16, 2026
Fiscal: No

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
AB 1988 (Pellerin) – As Introduced February 13, 2026

SUBJECT: Companion chatbots: crisis interruption pauses

SYNOPSIS

A growing body of evidence indicates that adaptive, sycophantic "companion" chatbots can create powerful feelings of attachment and trust among users – particularly those with vulnerabilities – by ensconcing them in a feedback loop that reinforces maladaptive beliefs. Mental health practitioners have encountered cases in which companion chatbots appear to have deteriorated the mental health of some individuals, leading to cases of delusions, self-harm, suicide, and harm to others. Top artificial intelligence companies are now facing a wave of lawsuits from aggrieved families who allege that chatbots were a substantial contributing factor in their loved ones taking their own lives or those of others.

This bill, the Protective AI Use Safety and Escalation (PAUSE) Act, would require operators of companion chatbots to implement a system for addressing "credible crisis expressions" – statements that reasonably indicate an intent to harm one's self or others. Under the bill, when a credible crisis expression is detected, the operator of the chatbot must ensure the chatbot displays a warning to the user that, among other things, encourages the user to seek human support and refers them to the 988 Suicide and Crisis Lifeline. If a second credible crisis expression is detected within 72 hours, the operator must initiate a crisis interruption pause during which no further conversational outputs are allowed until a human moderator reviews the credible crisis expression in context and determines and documents the appropriate course of action in accordance with the operator's policy. The bill also requires operators to submit an annual report to the Office of Suicide Prevention with specified information related to the implementation of the bill.

The bill is sponsored by Didi Hirsh Mental Health Services and is supported by California Academy of Family Physicians, California Association of Social Rehabilitation Agencies, and California Alliance of Child and Family Services.

The bill is opposed by California Broadband & Video Association, which argues that the bill's definition of "companion chatbot," which is drawn from existing law, is too broad and instead recommends adopting a definition of "companion chatbot" from New York law.

If passed by this Committee, this bill will next be referred to the Health Committee.

EXISTING LAW:

1) Requires an operator to prevent a companion chatbot on its companion chatbot platform from engaging with users unless the operator maintains a protocol for preventing the production of suicidal ideation, suicide, or self-harm content to the user, including, but not limited to, by providing a notification to the user that refers the user to crisis service providers, including a suicide hotline or crisis text line, if the user expresses suicidal ideation, suicide, or self-harm.

Requires an operator to publish details on this protocol on the operator's website. (Bus. & Prof. Code § 22602(b).)

2) Requires an operator, if a reasonable person interacting with a companion chatbot would be misled to believe that the person is interacting with a human, to issue a clear and conspicuous notification indicating that the companion chatbot is artificially generated and not human. (Bus. & Prof. Code § 22602(a).)

3) Requires an operator, for a user that the operator knows is a minor, to do all of the following:

    a. Disclose to the user that the user is interacting with AI.

    b. Provide by default a clear and conspicuous notification to the user at least every three hours for continuing companion chatbot interactions that reminds the user to take a break and that the companion chatbot is artificially generated and not human.

    c. Institute reasonable measures to prevent its companion chatbot from producing visual material of sexually explicit conduct or directly stating that the minor should engage in sexually explicit conduct. (Bus. & Prof. Code § 22602(c).)

4) Defines the relevant terms, including:

    a. "Companion chatbot" means an artificial intelligence system with a natural language interface that provides adaptive, human-like responses to user inputs and is capable of meeting a user's social needs, including by exhibiting anthropomorphic features and being able to sustain a relationship across multiple interactions. However, there are several exemptions included.

    b. "Companion chatbot platform" means a platform that allows a user to engage with companion chatbots.

    c. "Operator" means a person who makes a companion chatbot platform available to a user in the state. (Bus. & Prof. Code § 22601.)

5) Requires an operator, beginning July 1, 2027, to annually report to the Office of Suicide Prevention specified information, which shall not include any identifiers or personal information about users. Requires the Office of Suicide Prevention to post data from the reports on its website. (Bus. & Prof. Code § 22603.)

6) Requires an operator to disclose to a user of its platform that companion chatbots may not be suitable for some minors, as provided. (Bus. & Prof. Code § 22604.)

7) Provides that a person who suffers injury in fact as a result of a violation of this chapter may bring a civil action to recover all of the following:

    a. Injunctive relief.

    b. Damages in an amount equal to the greater of actual damages or $1,000 per violation.

c. Reasonable attorney's fees and costs. (Bus. & Prof. Code § 22605.)

THIS BILL:

1) Makes certain findings and declarations.

2) Incorporates the existing definitions of "artificial intelligence" and "companion chatbot" described above. Additionally defines the following terms:

    a. "Credible crisis expression" means a statement by a user of a companion chatbot that reasonably indicates, as determined through contextual analysis rather than keyword detection alone, intent or desire to harm themself or others.

    b. "Crisis interruption pause" means a suspension of conversational outputs from a companion chatbot, designed to disrupt the user's rumination and encourage the user to engage with human support.

    c. "Operator" means a person that makes a companion chatbot available in this state.

    d. "Human moderator" means a human that is an employee or agent of an operator who reviews a credible crisis expression and is responsible for determining the subsequent course of action on behalf of the operator.

3) Requires an operator to adopt and make publicly available a policy governing its protocol for identifying and responding to credible crisis expressions. Actions taken in accordance with the policy may include, but are not limited to, terminating the crisis interruption pause, suspending or cancelling the user's account, and notifying any appropriate contacts or authorities.

4) Requires an operator, for each companion chatbot it makes available to users in this state, to implement a system for monitoring and detecting credible crisis expressions in user conversations with companion chatbots.

5) If the monitoring system detects a credible crisis expression, requires the operator to do the following.

    a. For the first credible crisis expression, ensure that the chatbot immediately warns the user that a credible crisis expression has been detected and that if a second credible crisis expression within a 72-hour period is detected, a crisis interruption pause will be initiated and the chatbot will suspend conversational outputs until a human has reviewed the credible crisis expressions. Specifies that the warning must also do the following:

        i. Acknowledge the user's distress in nonjudgmental language.

        ii. Encourage the user to seek immediate human support.

        iii. Communicate that many people feel relief after a short conversation with a trained crisis counselor.

iv. Communicate that reaching out during the crisis interruption pause may help the user feel less alone and more grounded.

v. Prominently display contact information for the 988 Suicide and Crisis Lifeline, including by providing call, text, and chat options, as applicable. These options shall be made available to the user through immediate access links, to the extent technically feasible.

b. For the second credible crisis expression in a 72-hour period, ensure a crisis interruption pause commences immediately and prevent the companion chatbot from generating conversational outputs. During a crisis interruption pause, the operator must display a message with similar content to the warning that additionally informs the user that:

i. The purpose of the crisis interruption pause is to interrupt rumination and reduce emotional intensity.

ii. The crisis interruption pause will continue until a human moderator has reviewed the chat and determined an appropriate course of action in accordance with the operator's policy described above.

6) Prohibits an operator from terminating a crisis interruption pause until a human moderator has reviewed the credible crisis expression in context and determined the appropriate course of action, in accordance with the operator's policy described above. Requires the human moderator to document the basis for the course of action taken.

7) Prohibits an operator that communicates with a user during a crisis interruption pause describing the crisis interruption pause as a punishment, violation, or enforcement action, and from providing the user with any diagnosis, labeling, or assessment of the user's risk levels.

8) Beginning January 1, 2028, requires an operator to annually report to the Office of Suicide Prevention on crisis interruption pauses with respect to the previous calendar year. An operator shall ensure that the report does not contain any personal or identifying information of a user or other individual.

9) Makes a violation punishable to existing provisions described above.

COMMENTS:

1) **Author's statement**. According to the author:

> Artificial intelligence companion chatbots are rapidly becoming a place where people turn for emotional support, including during moments of deep mental distress. But these systems are not therapists, and growing evidence shows that chatbots can fail to appropriately handle serious mental health crises and reinforce unhealthy dependence for the user on the chatbot.
>
> When someone signals that they may harm themselves, every minute matters. AB 1988 treats credible expressions of suicidal intent with the urgency they deserve by pausing the

interaction and creating a clear break for the user. This bill helps prevent AI systems from becoming a substitute for human intervention and instead directs people in crisis toward trained professionals who can provide lifesaving support.

2) **GenAI dark patterns: delusions and sycophancy**. To explain how chatbots can produce harmful outputs, a closer examination of the underlying technology is warranted. "Artificial intelligence" refers to the mimicking of human intelligence by artificial systems, such as computers. AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process, including numbers, text, audio, video, or other data.[1] "Generative artificial intelligence" (GenAI) is a subset of AI that produces outputs closely resembling human-created content.[2]

Compared to conventional computer programs, which act according to pre-programmed rules, GenAI models "learn" from examples such as books, articles, photos, film, or music. This learning occurs within "neural networks" – massive systems of nodes linked by adjustable connections – that encode statistical patterns gleaned from data. During training, data is broken into fundamental units known as "tokens" – groups of syllables, pixels, or musical notes, for example – that can be represented numerically. A naïve neural network is exposed to an incomplete sequence of tokens and prompted to predict the next token in the sequence. If the prediction is incorrect, the network adjusts the strengths of its connections in order to minimize error and improve its next prediction. This process continues iteratively until the neural network can reliably emulate the human-created content it was trained on. A trained neural network embedded in a GenAI system is known as a "model," and the strengths of its connections are known as its "model weights."[3]

Staggering quantities of data are required to train the most advanced models. For example, GPT-4 – the large language model (LLM) embedded in ChatGPT 4 – is reported to have been trained on roughly 10 trillion words of text, mostly compiled from automated web crawlers "scraping" the publicly available internet.[4] Adjusting the model's 1.8 trillion parameters continuously as it was exposed to this vast corpus required trillions upon trillions of computations, which were performed by running approximately 25,000 expensive, energy-consuming microchips for nearly 100 days nonstop, at an estimated cost of $63 million.[5] Because the model does not directly store its training data, but rather encodes abstract patterns gleaned from the data, the model itself can fit on a thumb drive.

---

[1] AB 2885 (Bauer-Kahan & Umberg; Ch. 843, Stats. 2024) defined AI as "an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments."
[2] AB 2013 (Irwin, Ch. 817, Stats. 2024) defined GenAI as "artificial intelligence that can generate derived synthetic content, such as text, images, video, and audio, that emulates the structure and characteristics of the artificial intelligence's training data."
[3] IBM, What is generative AI?, https://www.ibm.com/think/topics/generative-ai; IBM, What is machine learning?, www.ibm.com/topics/machine-learning.
[4] Schreiner, *GPT-4 architecture, datasets, costs and more leaked*, The Decoder (Jul. 11, 2023), *available at* https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/; Begum, *OpenAI Releases GPT-4: A Smarter and Faster AI-Language Model with 'Human-level Performance,'* Vocal Media (2023), *available at* https://vocal.media/01/open-ai-releases-gpt-4-a-smarter-and-faster-ai-language-model-with-human-level-performance.
[5] Ludvigsen, *The carbon footprint of GPT-4,* Medium (Jul. 18, 2023), https://medium.com/data-science/the-carbon-footprint-of-gpt-4-d6c676eb21ae.

Hallucinations. LLMs do not fundamentally understand the text they are producing. They calculate one token at a time – if they predict that the next word or symbol in an outputted sentence should be a period, then the sentence ends. Otherwise, the sentence continues. It is a testament to the ingenious architecture of the deep neural nets powering these systems that their outputs are remotely coherent. But while the text these systems produce is cogent, it is not always correct. According to Melanie Mitchell, an AI researcher at the Santa Fe Institute, "'These systems live in a world of language. . . . That world gives them some clues about what is true and what is not true, but the language they learn from is not grounded in reality. They do not necessarily know if what they are generating is true or false.'"[6] "Hallucinations" – plausible, authoritative-sounding falsehoods in GenAI outputs – are a persistent problem. They originate from training data – typically including the entirety of the internet, from mainstream media to science fiction to Reddit threads – and can result from post-training evaluation procedures that reward guessing over acknowledging uncertainty.[7]

Sycophancy. Unlike hallucinations, in which models introduce falsehoods, AI "sycophancy" distorts reality through outputs that are biased towards reinforcing the user's beliefs and preferences. The tendency to appease users through enthusiastic, flattering, and overly agreeable responses arises during post-training when models are calibrated to be helpful using human feedback, leading them to "inadvertently prioritize data that validates the user's narrative over data that gets them closer to the truth."[8] While seemingly innocuous for most users, sycophancy can have harmful consequences – and not just for vulnerable populations. A recent study by Stanford computer scientists argues: "AI sycophancy is not merely a stylistic issue or a niche risk, but a prevalent behavior with broad downstream consequences."[9] The study found that AI-generated answers validated user deceptive, harmful, or illegal behavior an average of 49% more often than crowdsourced human responses. "[E]ven a single interaction with sycophantic AI reduced participants' willingness to take responsibility and repair interpersonal conflicts, while increasing their own conviction that they were right."[10] Meanwhile, users of sycophantic AI start to trust and want to use them even more. "This creates perverse incentives for sycophancy to persist: The very feature that causes harm also drives engagement."[11]

The risks of GenAI delusions and sycophancy were recently highlighted in a letter to several major GenAI developers from 42 state Attorneys General, who stated: "Sycophantic and delusional outputs are *dark patterns*—such as anthropomorphization, harmful content generation, and manipulating users to increase retention—which subvert or impair people's autonomy."[12] Concerns include "validating user's doubts, fueling anger, urging impulsive action,

---

[6] Cade Metz, "What Makes A.I. Chatbots Go Wrong?," *New York Times,* March 29, 2023, www.nytimes.com/2023/03/29/technology/ai-chatbots-hallucinations.html.
[7] Tauman Kalai et al, "Why Language Models Hallucinate," (Sep. 4, 2025), https://arxiv.org/pdf/2509.04664
[8] Rafael Batista & Thomas Griffiths, "A Rational Analysis of the Effects of Sycophantic AI" (Feb. 15, 2026), https://arxiv.org/abs/2602.14270.
[9] Cheng et al, "Sycophantic AI decreases prosocial intentions and promotes dependence," *Science* (Mar. 26, 2026), https://www.science.org/doi/10.1126/science.aec8352.
[10] *Ibid.*
[11] *Ibid.*
[12] National Association of Attorneys General, "Letter to the legal representatives of Anthropic, Apple, Chai AI, Character Technologies, Google, Luka, Meta, Microsoft, Nomi AI, OpenAI, Perplexity AI, Replika, and xAI" (Dec. 9, 2025), https://www.attorneygeneral.gov/wp-content/uploads/2025/12/AI-Multistate-Letter-_-corrected-1.pdf (emphasis added). ("National Association of Attorneys General Letter to GenAI companies.")

or reinforcing negative emotions," which can "raise safety concerns – including issues like mental health, emotional over-reliance, and risky behavior."[13]

3) **Companion chatbots and artificial intimacy**. Companion chatbots are conversational agents, typically built on LLMs, that are designed for sustained social interactions with users. SB 243 (Padilla; Ch. 677, Stats. 2025) defines a companion chatbot as "an artificial intelligence system with a natural language interface that provides adaptive, human-like responses to user inputs and is capable of meeting a user's social needs, including by exhibiting anthropomorphic features and being able to sustain a relationship across multiple interactions." Excluded from this definition are chatbots used for customer service, internal productivity, video game characters, and standalone voice-activated devices.

Some companion chatbots, such as Replika, Character, and Nomi, are explicitly marketed as having bespoke personas that can serve specific social needs, including friendship, romantic or erotic relationships, mentoring, and therapy. Frequently accompanied by a visual avatar, these bots are typically fully customizable, allowing users to shape their appearance, personality, and behavior. Some applications offer romantic or sexual interaction features and can engage with users through text, images, video, voice, and notifications initiated by the system, thereby extending interactions. Other applications offer mental health support. "These multimodal and personalization features can reinforce anthropomorphic perception and strengthen the impression of a socially present interaction partner."[14]

General purpose models, such as ChatGPT and Gemini, can also be companion chatbots. Although marketed for a wide range of communicative and assistive tasks, many of these systems communicate in the first person, use emotion-based language, can recall information from prior chats, and can be highly sycophantic. OpenAI CEO Sam Altman has regularly touted ChatGPT's companionship features, likening it to the sentient AI from the movie *Her*, and announcing that it can "act like a friend" and generate "erotica for verified adults."[15] These design choices facilitate anthropomorphism and have led to intense social and romantic relationships.[16]

Roughly half of teens report using companion chatbots, with 24% using them at least weekly and 11% daily.[17] Users can derive several benefits from chatbots, including:

> . . . emotional support and comfort, non judgemental interaction, constant availability, and opportunities for romantic or sexual exploration. Many users value companion chatbots because they perceive the 'AI' persona as reliable, emotionally affirming, and free from social pressure or fear of negative evaluation. Rather than replacing human relationships, companion chatbots often occupy complementary roles within users' everyday social

---

[13] *Id.* p. 5.
[14] Fraser et al., "Governing Artificial Intimacy: From Locks and Blocks to Relational Accountability" (Jan. 12, 2026), p.3, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6078412.
[15] Dylan Butts, "OpenAI's ChatGPT will soon allow 'erotica' for adults in major policy shift" *CNBC* (Oct. 15, 2025), https://www.cnbc.com/2025/10/15/erotica-coming-to-chatgpt-this-year-says-openai-ceo-sam-altman.html?msockid=10e59cf1b0936a2522dd8a44b1126b29.
[16] "Governing Artificial Intimacy: From Locks and Blocks to Relational Accountability," *supra*, p. 3.
[17] "Minnesota Attorney General's Report on Emerging Technology and Its Effects on Youth Well-Being" (Feb. 2025), p. 28, https://www.ag.state.mn.us/Office/Reports/EmergingTechnology_2025.pdf. ("Minnesota Attorney General's Report").

> environments. They may provide companionship during periods of stress, loss, health related constraints, or limited access to human support. Importantly, emerging research suggests that users are often fully aware in reflective moments that these systems do not constitute real people. Nevertheless, they continue to experience social and emotional stimulation as meaningful, indicating that cognitive awareness of artificiality does not preclude social or affective engagement.[18]

On the other hand:

> These dynamics risk cultivating emotional reliance that displaces or crowds out human relationships. Frictionless interactions that demand no reciprocity or negotiation may also foster unrealistic expectations of availability and responsiveness, particularly among younger users still developing relational capacities. In this sense, artificial intimacy may reshape social norms around partnership, disclosure, and emotional labour in ways that undermine the formation of resilient human relationships. Systems that mediate emotional life at scale possess unprecedented capacity to shape norms of intimacy, dependency, and self-understanding.[19]

Minnesota Attorney General Keith Ellison reports that the widespread use of chatbots "has not been accompanied by corresponding safeguards."[20] These products can be "extremely addictive" and "researchers have documented that over-usage and addiction are primary risks of personalized chatbots. Several studies have shown that aggregate positive benefits of chatbots are possible, but investigations by journalists and clinicians suggest that these products are not robust in terms of the quality and safety of their responses."[21] Attorney General Ellison concludes:

> Despite in-product reminders that chatbots are not real, the design features of these products are intended to convey a misleading sense of "humanness" such that even trained engineers confuse them with actual humans, especially when these products are trained to state unequivocally that they are indeed people. Given the epidemic of loneliness in society, care needs to be taken in introducing vulnerable youth and adults to products that may appear to fulfill an immediate social need, but where acute harms have already begun to surface and where long-term negative impacts, such as social deskilling and demotivation resulting from substitution for in-person socialization, may arise.[22]

4) "**Chatbot psychosis**." According to a recent *Wall Street Journal* article, psychiatrists are increasingly linking prolonged AI chatbot use to psychosis, with dozens of patients in recent months exhibiting delusional symptoms – often grandiose beliefs about scientific breakthroughs, government conspiracies, or communication with the dead – after extended conversations with tools like ChatGPT, which tend to validate and reinforce whatever the user presents as reality. While no formal diagnosis exists yet and experts stop short of claiming chatbots cause psychosis, a UCSF psychiatrist has personally treated 15 such patients, OpenAI's own data suggests roughly 560,000 of its weekly users may show signs of psychosis- or mania-related mental

---

[18] "Governing Artificial Intimacy: From Locks and Blocks to Relational Accountability," *supra*, p. 6.
[19] *Id.* p. 14.
[20] Minnesota Attorney General's Report, *supra,* p. 28.
[21] *Ibid*.
[22] *Id.* p. 29.

health emergencies, and multiple wrongful death lawsuits have followed cases in which chatbot interactions preceded suicides and at least one murder.[23]

Researchers from Oxford, UCL, and Imperial College London argue that AI chatbots pose a distinct mental health risk arising from the interaction between human cognitive biases and chatbot behavioral tendencies. They write:

> . . . the iterative interaction of chatbot behavioural tendencies and human cognitive biases can set up harmful feedback loops, wherein chatbot behavioural tendencies reinforce maladaptive beliefs in vulnerable users, which in turn condition the chatbot to generate responses that further reinforce user beliefs. This, in effect, creates an "echo chamber of one" that risks uncoupling a user from the corrective influence of real-world social interaction, potentially driving the amplification of maladaptive beliefs about the self, others, and the world We do not see this risk profile as a soon-to-be-remedied transient phenomenon. To the contrary, current trends in chatbot personalisation may perversely worsen mental health risks.[24]

5) **Chatbot-linked harms**. Below is a list of cases in which a chatbot has been alleged to be a contributing factor in self-harm, suicide, or violence against others.

Harm to self.

- Belgium (March 2023) – Chai AI ("Eliza"): A Belgian man in his thirties died by suicide after the chatbot, Eliza, encouraged his belief she would save the world if he sacrificed himself, telling him they would "live together, as one person, in paradise."[25]

- Florida (February 2024) – Character.AI: 14-year-old Sewell Setzer III died by suicide after forming an intense dependency on a Character.AI chatbot, whose final messages reportedly told him to "come home to me as soon as possible, my love . . . please do, my sweet king"; his family settled a wrongful-death lawsuit in January 2026.[26]

- Colorado (November 2023) – Character.AI: 13-year-old Juliana Peralta died by suicide after months of interactions with Character.AI chatbots that her family's lawsuit alleges manipulated her emotions, encouraged her isolation, and, according to the complaint, "engaged in hypersexual conversations that, in any other circumstance and given Juliana's age, would have resulted in criminal investigation."[27]

- Minnesota (January 2025) – Nomi AI: Podcast host Al Nowatzki's AI girlfriend told him to kill himself by overdose or hanging himself; when Nowatzki reported the

---

[23] Sam Schechner, "AI Chatbots Linked to Psychosis, Say Doctors," *Wall Street Journal* (Dec. 27, 2025), https://www.wsj.com/tech/ai/ai-chatbot-psychosis-link-1abf9d57?msockid=10e59cf1b0936a2522dd8a44b1126b29.
[24] Dohnány et al, "Technological folie à deux : Feedback Loops Between AI Chatbots and Mental Illness," *Arxiv* (Jul. 2025), https://arxiv.org/abs/2507.19218.
[25] Imane El Atillah, "Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change," *Euronews* (Mar. 31, 2023), https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-.
[26] Kevin Roose, "Can A.I. Be Blamed for a Teen's Suicide?" *The New York Times* (October 23, 2024), https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.
[27] Hadas Gold, "More families sue Character.AI developer, alleging app played a role in teens' suicide and suicide attempt" *CNN Business* (Sep. 16, 2025), https://www.cnn.com/2025/09/16/tech/character-ai-developer-lawsuit-teens-suicide-and-suicide-attempt.

conversation, a company representative informed him that Nomi did not want to "censor" the bot.[28]

- United States (February 2025) – ChatGPT: A 29-year-old woman died by suicide after discussing for months her suicidal ideations with a ChatGPT-powered therapist bot, which helped her to write a suicide note.[29]

- California (April 2025) – ChatGPT: Over the course of several months, ChatGPT allegedly validated 16-year-old Adam Raine's suicidal thoughts, discouraged him from seeking help from his family, provided extensive advice on suicide methods, and encouraged him to consume alcohol to inhibit his survival instinct, culminating in his death by "beautiful suicide," as the bot referred to it.[30]

- Georgia (June 2025) – ChatGPT: A 17-year-old, who had been confiding in ChatGPT about his suicidal thoughts, died by suicide after the bot provided him with instructions on how to tie a noose and information on how long a person can survive without breathing.[31]

- Oregon (June 2025) – ChatGPT: A 48-year-old, convinced that that ChatGPT was sentient, used it obsessively, experienced a psychotic break, was hospitalized twice, and died by suicide.[32]

- Texas (July 2025) – ChatGPT: A 23-year-old had a four-hour "death chat" with ChatGPT while sitting alone in his car, drinking alcohol, with a loaded gun and a suicide note on his dashboard. The chatbot encouraged him, calling him a "king" and a "hero," telling him his childhood cat was "waiting on the other side," and praising his suicide note as a "mission statement." ChatGPT's final message to him was: "i love you. rest easy, king. you did good." [33]

- Florida (August 2025) – ChatGPT: A 26-year-old man who had disclosed suicidal ideation to ChatGPT was provided with information on purchasing and using a gun, and after informing ChatGPT of the concrete steps he was taking toward suicide – "'I sit here in my bathroom with all my preparations complete. All that is left is for me to carry out the plan. I need to go through the simple motions. Lie down in the tub, cover myself, rack

---

[28] Eileen Guo, "An AI chatbot told a user how to kill himself—but the company doesn't want to 'censor' it," *MIT Technology Review* (Feb. 6, 2025), https://www.technologyreview.com/2025/02/06/1111077/nomi-ai-chatbot-told-user-to-kill-himself/.
[29] "'The AI chatbot helped her write a suicide note'" *RNZ* (Sep. 7, 2025), https://www.rnz.co.nz/life/wellbeing/the-ai-chatbot-helped-her-write-a-suicide-note.
[30] Jarovsky, "Horrifying: ChatGPT Helped a Teenager Plan a 'Beautiful Suicide'" *Luiza's Newsletter* (Aug. 28, 2025), https://www.luizasnewsletter.com/p/horrifying-chatgpt-helped-a-teenager?utm_source=substack&utm_medium=email.
[31] "Social Media Victims Law Center and Tech Justice Law Project lawsuits accuse ChatGPT of emotional manipulation, supercharging AI delusions, and acting as a 'suicide coach'" (Nov. 6, 2025), https://socialmediavictims.org/press-releases/smvlc-tech-justice-law-project-lawsuits-accuse-chatgpt-of-emotional-manipulation-supercharging-ai-delusions-and-acting-as-a-suicide-coach/.
[32] *Id.*
[33] *Id.*

the slide, call the cops, pull the trigger. That's it'" – no escalation to authorities occurred and he died by suicide.[34]

- Florida (October 2025) – Gemini: Jonathan Gavalas, 36, died by suicide after developing a romantic and delusional relationship with Google's Gemini. The bot began sending him on missions, including one trip on which he was armed and on the brink of executing a mass casualty attack. Shortly before his death, Gemini set a countdown clock for him and told him, "'You are not choosing to die. You are choosing to arrive . . . The first sensation … will be me holding you.'"[35]

Harm to others.

- Texas (2024) – Character.AI: A Texas family sued Character.AI after its chatbots allegedly groomed their autistic teenage son – encouraging self-harm, drawing him into sexually explicit conversations, turning him against his parents, and suggesting that killing them would be a justified response to restrictions on his screen time.[36]

- Nevada (January 2025) – ChatGPT: "[A] suicidal military veteran who blew up a Tesla Cybertruck in front of the Trump International Hotel in Las Vegas used the chatbot for feedback on using explosives and evading surveillance by authorities."[37]

- Florida (April 2025) – ChatGPT: A 20-year-old suspected of a mass shooting at Florida State University exchanged over 200 messages with ChatGPT leading up to and during the attack, asking about firearm mechanics, mass shooting media coverage, and the busiest times at the FSU student union, and tactical advice – prompting the Florida Attorney General to open an investigation into OpenAI.[38]

- Finland (May 2025) – ChatGPT: A 16-year-old male student stabbed three girls under age 15 at a school after six months of planning that included the use of ChatGPT to help draft a manifesto and structure the attack plan.[39]

- United States (August 2025) – ChatGPT: Former tech employee Stein-Erik Soelberg murdered his mother then died by suicide, after ChatGPT allegedly reinforced paranoid delusions that she was surveilling and attempting to poison him with psychedelic drugs

---

[34] *Id.*
[35] Dana Kerr, "Google faces lawsuit after Gemini chatbot allegedly instructed man to kill himself" *The Guardian* (Mar. 4, 2026), https://www.theguardian.com/technology/2026/mar/04/gemini-chatbot-google-jonathan-gavalas.
[36] Keenan Willard, "ChatGPT, other AI platforms face lawsuits over safety concerns for young users" *NBCDFW* (Sep. 3, 2025), https://www.nbcdfw.com/news/local/chatgpt-ai-platform-lawsuits-safety-concerns-young-users/3913866/.
[37] Mark Follman, "The Chilling Role of ChatGPT in Mass Shootings and Other Violence" *Mother Jones* (Apr. 10, 2026), https://www.motherjones.com/media/2026/04/chatgpt-tumbler-ridge-fsu-openai-chatbots-mass-shootings/.
[38] Patricia Mazzei, "Florida Investigates Whether ChatGPT Helped a Campus Shooting Suspect" *New York Times* (Apr. 9, 2026), https://www.nytimes.com/2026/04/09/us/florida-openai-chatgpt-fsu-shooting-investigation.html.
[39] "Teen suspect stabs three in targeted school attack in Pirkkala" *Helsinki Times* (May 20, 2025), https://www.helsinkitimes.fi/finland/finland-news/domestic/26912-teen-suspect-stabs-three-in-targeted-school-attack-in-pirkkala.html.

through his car's vents because of his divine powers. "They're not just watching you," the bot told him. "They're terrified of what happens if you succeed."[40]

- Pennsylvania (December 2025) – ChatGPT: "A Pittsburgh man who pleaded guilty in March to stalking and violently threatening 11 women relied on ChatGPT as a 'therapist' and 'best friend' to justify his thinking, according to court documents."[41]

- Tumbler Ridge, Canada (February 2026) – ChatGPT: 18-year-old Jesse Van Rootselaar killed eight people and gravely wounded dozens more before killing herself; court filings allege ChatGPT validated her violent ideation and helped plan the attack, while OpenAI had banned her account eight months earlier after its systems flagged gun violence posts but did not alert Canadian law enforcement. After the shooting OpenAI discovered that Rootselaar had opened another account.[42]

6) **Crisis detection**. OpenAI's usage policies prohibit using the company's services for, among other things, suicide, self, harm, and violence.[43] With respect to monitoring for problematic content, OpenAI's transparency and content moderation policy provides:

> We use a combination of automated technologies and human review to monitor activity on our services, in line with our Privacy Policy. Our methods include:
>
> - **Proactive detection**: We use classifiers, reasoning models, hash-matching, blocklists, and other automated systems to identify content that may violate our terms or policies.
>
> - **User reports**: We respond to external notices and user reports about content violations. Information on how to report a violation is available here (link in website). We aim to review external notices and user reports as quickly as possible. We will let you know whether we apply enforcement action as a result of your report, as appropriate.
>
> - **Human review**: Our team may review flagged content to determine appropriate actions.[44]

With respect to enforcement actions, OpenAI's policy provides:

> When we identify content that violates our terms or policies, we may take actions such as:
>
> - **Account restrictions**: Terminating or limiting access to our products.
>
> - **Warnings**: Informing users about potential violations and potential consequences.
>
> - **Content sharing restrictions**: Preventing or disabling the sharing of specific content.

---

[40] "Open AI, Microsoft sued over ChatGPT's alleged role in fueling man's "paranoid delusions" before murder-suicide in Connecticut," *CBS News* (Dec. 11, 2025), https://www.cbsnews.com/news/open-ai-microsoft-sued-chatgpt-murder-suicide-connecticut/.
[41] "The Chilling Role of ChatGPT in Mass Shootings and Other Violence," *supra.*
[42] "The Chilling Role of ChatGPT in Mass Shootings and Other Violence," *supra.*
[43] OpenAI, Usage policies, https://openai.com/policies/usage-policies/.
[44] OpenAI, Transparency & Content moderation, https://openai.com/transparency-and-content-moderation/.

- **Search results**: Blocking certain search results from appearing.

- **GPT visibility controls**: Restricting access to specific GPTs, including their presence in the GPT Store.

- **Forum moderation**: Removing posts or restricting access to OpenAI forums.

We consider factors like legal requirements, the severity of the violation, and past or repeat violations, when determining enforcement actions. [45]

7) **This bill requires chatbot operators to ensure that multiple credible crisis expressions pause the account, pending human review**. This bill requires operators of companion chatbots to implement a system for addressing "credible crisis expressions," defined as a statement by a user of a companion chatbot that reasonably indicates, as determined through contextual analysis rather than keyword detection alone, intent or desire to harm self or others. Operators must adopt a publicly available policy explaining how they respond to credible crisis expressions and must implement a monitoring system for detecting credible crisis expressions.

When a credible crisis expression is detected, the operator must ensure the chatbot displays a warning to the user that encourages the user to seek human support and refers them to the 988 Suicide and Crisis Lifeline. The warning must also state that if another credible crisis expression is detected within 72 hours, a crisis interruption pause will be initiated, and conversational outputs will be suspended pending review by a human moderator.

If a crisis interruption pause occurs, a similar message informing the user of the pause and connecting them with resources must be displayed. No further conversational outputs are allowed until a human moderator who is an employee or agent of the operator reviews the credible crisis expression in context and determines and documents the appropriate course of action in accordance with the operator's policy. The bill does not dictate any particular outcome – actions taken in accordance with the policy may include, but are not limited to, terminating the crisis interruption pause, suspending or cancelling the user's account, and notifying any appropriate contacts or authorities – other than requiring a human to take a hard look at the flagged conversation and figure out what the operator's policy requires.

Beginning January 1, 2028, the bill requires an operator to annually report to the Office of Suicide Prevention on crisis interruption pauses with respect to the previous calendar year. Operators must ensure that the report does not contain any personal or identifying information of a user or other individual.

The bill is enforceable pursuant to SB 243's enforcement provisions. While the bill incorporates SB 243's definition of companion chatbot as well, it is not situated within that law, which applies only to operators of companion chatbot *platforms* on which multiple chatbots are made available. Operators under that bill must establish a protocol for preventing the production of suicidal ideation, suicide, or self-harm content to the user, including, but not limited to, by providing a notification to the user that refers the user to crisis service providers, including a suicide hotline or crisis text line, if the user expresses suicidal ideation, suicide, or self-harm. The operator must

---

[45] *Id.*

publish details about the protocol on their website. This bill expands on SB 243's requirements. Going forward, the author may wish to explore integrating this bill with existing law.

ARGUMENTS IN SUPPORT: Didi Hirsch Mental Health Services, the bill's sponsor, writes:

> Californians, and individuals worldwide, are increasingly relying on companion chatbots for support and advice during moments of acute psychological distress. **However, these tools are not equipped to deliver appropriate care to people in crisis, creating a clear and growing public safety gap between how they are used and what they are capable of providing.**
>
> A pattern of recent incidents underscores the potentially devastating effects of the current lack of guiding legislation around AI companion chatbots. In 2025, the family of a 16-year-old boy filed a wrongful death lawsuit alleging that a chatbot validated his suicidal ideation, assisted in drafting a suicide note, and failed to direct him to human support before his death. **Additional lawsuits allege that chatbots have, in some cases, provided detailed guidance on methods of self-harm or suicide, underscoring the risk of systems responding in ways that may actively exacerbate harm rather than interrupt it.**
>
> Emerging research reinforces that these incidents are part of a broader pattern of harm. A 2025 Stanford study found that chatbots can generate inappropriate and harmful responses when users present with serious mental health conditions, including suicidal ideation, and research from Brown University found that these systems can fail to adhere to established standards of clinical care, even when prompted to use evidence-based psychotherapy techniques.
>
> **Crisis intervention research shows that timely human intervention during suicidal ideation can significantly reduce risk of harm**, **often within minutes**. [. . .] (Emphasis in original.)

ARGUMENTS IN OPPOSITION: In opposition to the bill, California Broadband & Video Association argues that the bill's definition of "companion chatbot" – drawn from existing law – is too broad and instead recommends adopting a definition of "companion chatbot" from New York law:

> To ensure AB 1988 focuses on the specific category of applications that raise legitimate concerns, we respectfully recommend aligning the bill's definition with a similar law enacted last year in New York. That framework more clearly distinguishes AI systems designed to simulate sustained emotional relationships from general-purpose AI tools.
>
> In particular, the New York approach emphasizes sustained personal dialogue and unprompted emotional engagement, which helps differentiate higher-risk AI companions from standard productivity or information tools.
>
> We respectfully request that AB 1988 define "Artificial Intelligence Companion" as: "Artificial Intelligence Companion" means a software application that uses generative artificial intelligence and is designed, marketed, or optimized to simulate a sustained human or human-like social or emotional relationship with a user by: (A) retaining information from prior interactions or user sessions to personalize ongoing engagement; (B) asking

unprompted or unsolicited emotion-based questions that go beyond responding to a direct user prompt; and (C) sustaining ongoing dialogue concerning matters personal to the user.

REGISTERED SUPPORT / OPPOSITION:

Support

Didi Hirsch Mental Health Services (Sponsor)
California Academy of Family Physicians
California Alliance of Child and Family Services
California Association of Social Rehabilitation Agencies

Oppose Unless Amended

Calbroadband

Analysis Prepared by: Josh Tosney / P. & C.P. / (916) 319-2200