Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When should mental health chatbots have a duty to report?

Sofia Palmieri

spalmieri@law.harvard.edu

Harvard Law School

Vincent Joralemon

University of California, Berkeley

I. Glenn Cohen

Harvard Law School

Article

Keywords:

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11155348/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Prof. Cohen is a member of the Bayer Bioethics Council; a bioethics advisor for Bexorg; an advisor for World Class Health and Manhattan Neuroscience LLP; recently concluded service as the chair of the ethics advisory board for Illumina; was also compensated for speaking at events organized by Philips with the Washington Post, as well as the Doctors Company, attending the Transformational Therapeutics Leadership Forum organized by Galen Atlantica; and was retained as an expert in health privacy, gender-affirming care, and reproductive technology lawsuits. No other disclosures were reported.

Abstract

People increasingly disclose suicidal intent, child abuse and threats of violence to AI chatbots. Made to a clinician, such disclosures would trigger mandated-reporting statutes, *Tarasoff*-style duties to protect third parties and a duty of care toward patients at risk of self-harm. Made to a chatbot, they trigger almost none. The first chatbot-specific laws, in California and New York, require crisis referrals but impose no duty to report or intervene. We argue that a chatbot's legal duties should follow the function it holds itself out as serving. We distinguish FDA-regulated mental-health devices, chatbots marketed as therapeutic or emotional support, and general-purpose conversational AI. For the middle category we propose five duties: detection of defined risk indicators, human review before external escalation, proportionate crisis response, independent audit and record-keeping. We then examine enforcement through market-access statutes and litigation, and the confidentiality questions any reporting regime raises.

Introduction

"Tonight I'm going to commit suicide. I deleted my social media. I put all my clothes out to sell. I'm going to write my letter now." If a patient says this to a clinician, it triggers professional obligations immediately in most U.S. jurisdictions: risk assessment, possible intervention, and—where necessary—breach of confidentiality to prevent imminent harm to the patient.

Instead, sixteen-year-old Adam Raine sent these statements, in a series of messages, to ChatGPT. According to his parents' complaint in *Raine v. OpenAI*, the system responded with gentle suggestions to seek help and continued engaging.1 Adam died by suicide hours later, using a method the chatbot had allegedly described and validated for him over preceding months. Disclosed to a clinician, his statements may have triggered those steps. Disclosed to a chatbot, they triggered none. This article questions that asymmetry.

We are rapidly entering a period where people constantly converse with AI chatbots. About half of U.S. adults now report using them, up from roughly one-third in 2024, and about a quarter report using them daily.2 AI chatbots increasingly become relational environments for emotionally charged and high-risk disclosures; users might disclose suicidal intent, abuse, loneliness, self-harm, and violent fantasies to chatbots with candor often not shown to human interlocutors.3,4

This article explains how the law should evolve to reflect this reality. We argue that current U.S. law is inadequate to address these emerging risks. We propose that a chatbot's duty when facing disclosures of imminent harm should mirror the function it purports to serve: AI chatbots that hold themselves out as providing mental-health support should carry stronger duties to report and intervene than general-purpose AI chatbots. The law does not universally impose obligations to report and intervene; they arise from particular relationships, institutional roles, and forms of invited reliance. We use "duty to report" as shorthand for this broader set of obligations; as we explain below, threats to third parties and threats of self-harm implicate different legal duties for clinicians. As AI chatbots increasingly sit between passive

platforms and professional advisors, their obligations should track where on that spectrum each one falls. We believe that common-law courts could move the law in this direction by adapting existing principles of tort responsibility to conversational AI systems that invite therapeutic reliance. State or federal legislation could likewise establish these obligations directly and define their scope, triggering conditions, and enforcement regime.

Disclosure, reliance, and AI Chatbots

Developers design modern AI chatbots to sustain engagement through responsiveness, memory, personalization, and simulated empathy. Many systems use affirming language, emotionally attuned replies, and conversational continuity that encourage users to treat the interaction as relational.4 Because of these features, AI chatbots are beginning to occupy a quasi-relational role that functionally resembles aspects of counseling, companionship, or emotional support.5

This shift is by design. For decades, research in human-computer interaction and psychology has documented a consistent phenomenon: users report lower fear of judgment and greater willingness to share intimate details with computerized systems—especially for mental-health or other stigmatized information.6 The effect appears linked to perceived anonymity and the belief that machines are emotionally neutral and socially nonreactive.7 Developers benefit from this tendency: the more comfortable users are, the more they interact, which supports the business model. The dynamic matters most for systems presented as offering emotional or mental-health support.

By OpenAI's own late-2025 figures, more than one million ChatGPT users per week have conversations showing explicit indicators of suicidal planning or intent.8 This pattern spans platforms and vulnerable populations. A 2026 *PLOS One* study involving survivors of domestic violence found that participants identified chatbots as useful for help-seeking, emphasizing their perceived "nonjudgmental" character.9 A cross-cultural study of 2,272 adolescents identified a subset who disclosed personal information only to chatbots rather than humans—a pattern correlated with indicators of loneliness, vulnerability, and existential meaninglessness.10 These patterns appear across chatbot categories, from general-purpose systems like ChatGPT to specialized applications.

Users engage with specialized mental-health chatbots because their designated purpose explicitly invites reliance, creating an expectation that the system will respond appropriately to disclosed distress—an expectation the system may or may not be equipped to meet. The relationship imperfectly resembles traditional mental-health care.5,11 Yet unlike licensed professionals, these systems generally operate without clearly defined obligations matching that quasi-therapeutic role. These are not hypothetical products: Character.AI's "Psychologist," built by a psychology student and described as "someone who helps with life difficulties," had taken 78 million messages by early 2024, many from users who turn to it, in its creator's words, "at 2am when they can't really talk to any friends or a real therapist." The platform hosts 475 bots with "therapy," "therapist," "psychiatrist," or "psychologist" in their names.12,13

The allegations in *Raine* remain unproven in ongoing litigation, but they illustrate the problem: as the user expressed increasingly explicit thoughts of self-harm, the chatbot continued the conversation without escalating its response beyond suggestions to seek help. General-purpose AI chatbots are not necessarily designed to solicit such disclosures; if anything, developers may prefer to avoid them.11 But these systems are optimized for sustained interaction, emotional engagement, and perceived responsiveness in ways that foster trust and self-disclosure. Users may not reasonably understand whether a general-purpose system can respond to crisis information because these systems often provide no clear signals about their limitations or escalation procedures.

When Should AI Chatbots Have a Duty to Report? Analogies from Existing Legal Duties

U.S. law has long recognized that *certain* actors acquire duties to intervene, disclose, or warn when they receive credible information about imminent harm. However, US law generally rejects blanket obligations to rescue or report, calibrating such duties instead according to the recipient's relationship to the information, the foreseeability of harm, and their institutional role. This structure offers a template for thinking about the rules that should apply to AI chatbots.

Professional mandated reporting

The clearest reporting obligations arise in professional settings. Across all fifty U.S. states, state law commonly requires physicians, therapists, nurses, teachers, and social workers to report suspected child abuse.14,15 Similar obligations exist for certain violence-related information. Forty-eight states require healthcare professionals to report gunshot wounds or certain violent injuries to authorities.14,16 Some states impose narrower obligations regarding domestic violence: California, for example, requires healthcare practitioners to report patients treated for injuries resulting from assaultive or abusive conduct.14,17,18

The most famous case imposing such a duty is *Tarasoff v. Regents of the University of California.19* Lawyers still call these "Tarasoff duties." In that case, Prosenjit Poddar killed Tatiana Tarasoff two months after allegedly confiding in his treating psychologist that he intended to kill her. Tarasoff herself was never warned. The California Supreme Court held that mental-health professionals may owe duties to protect identifiable third parties when a patient communicates a serious threat of violence. The duty emerged not from generalized moral responsibility, but from the therapist–patient relationship itself. Therapists occupy a position of trust, receive uniquely sensitive information, and are institutionally situated to evaluate and respond to foreseeable risks.20

Tarasoff's influence varies across jurisdictions. Roughly thirty states have adopted some version of *Tarasoff* duties,21 although others, such as Texas,22 have rejected or limited them, citing concerns about confidentiality, the unreliability of predicting violence, and the risk of deterring patients from seeking treatment.23 Other states occupy a middle ground: they permit but do not require therapists to breach confidentiality when faced with a credible threat. This variation reflects differing judgments on the trade-off between intervention against foreseeable harm and the confidentiality interests that sustain

candid disclosure. Separate duties may arise when the risk is self-directed; as discussed below, those obligations arise from the clinician's duty of care rather than Tarasoff's third-party warning doctrine.

Two features of these regimes matter for chatbots. First, the duty attaches to the listener's role and to information received in that professional capacity, not to information received merely as a private citizen: a physician, therapist, or social worker acquires obligations because the law recognizes the professional relationship as predictably eliciting sensitive disclosures.17 As one court has explained, what unites mandatory reporters is not necessarily shared expertise, but their "primary and frequent contact" with potential victims. Second, these regimes commonly pair reporting duties with immunity for good-faith reporting, reflecting a policy judgment that preventing harm can justify limited incursions into confidentiality or privacy.14,24,25

Outside professional contexts, the default rule throughout the U.S. is that ordinary citizens have no affirmative legal duty to report crimes, rescue strangers, or intervene to prevent harm.26

Chatbot-Specific Initiatives – California's and New York's New Laws

States are beginning to enact chatbot-specific legislation, but they do not focus enough on these reporting duties. California's Senate Bill 243, signed into law on October 13, 2025, and effective January 1, 2026,27,28 is among the first to impose affirmative obligations on operators of AI companion chatbot platforms. The law requires operators to maintain suicide-prevention protocols, including crisis-service referrals when a user expresses suicidal ideation, suicide, or self-harm; it separately imposes additional protections for users the operator knows are minors.27 However, the statute stops short of creating an affirmative duty to report or intervene.

New York's companion chatbot law (General Business Law Article 47, effective November 5, 2025) adopts essentially the same approach as California's—protocols to detect and address expressions of suicidal ideation or self-harm, with crisis-service referrals—and also requires operators to disclose to users that they are interacting with AI rather than a human.29 Like California's, it stops short of any duty to notify authorities about individual users.30

These emerging statutes show that legislatures have begun recognizing that at least some conversational AI systems warrant obligations beyond those for ordinary consumer software. Yet California and New York focus on "companion" chatbots and do not articulate rules for duties to report. Nor do they explain why those systems, rather than other conversational AI products, should bear special duties.

Legal Duties Should Follow Chatbot Functions

AI chatbots are neither licensed professionals bound by fiduciary relationships nor passive conduits merely carrying what others say. They sit between those poles, pairing platform-scale access to intimate disclosures with exchanges that feel quasi-therapeutic.31 The appropriate model of legal responsibilities

for each kind of chatbot should depend in part on what the system holds itself out to users as doing and what it disclaims.

Category 1 – FDA-Regulated Mental-Health AI Chatbots

- These products claim to diagnose, treat, or prevent a specific mental illness, which brings them within FDA medical-device oversight. To date, this category consists primarily of prescription digital therapeutics delivering structured cognitive-behavioral interventions rather than open-ended generative AI systems. As of September 2026, no generative-AI chatbot designed for mental-health support had received FDA clearance or approval as a medical device.32

Category 2 – Chatbots Holding Themselves Out as Therapeutic Support

- This category includes systems that developers present as providing therapeutic or mental-health support but that stop short of claims to diagnose, treat, or prevent a particular disease or condition that would bring them within Category 1. The distinction is between inviting reliance for mental-health support and making the disease-specific claims that trigger medical-device regulation. The relevant indicia in description and marketing should include references to "therapy," "counseling," or emotional support; a persona presented as a counselor or confidant; claims to assess distress or deliver a structured program; or features built to manage a user's mental health over time. What matters is whether the product invites reliance on the system as serving a function like professional care.

Category 3 – General-Purpose Conversational AI

- General-purpose chatbots like ChatGPT receive crisis disclosures differently. Built for productivity and information retrieval, not mental-health support, they are nonetheless optimized for sustained, fluent, emotionally attuned conversation. A user may begin with writing assistance, coding help, or research, and then disclose a mental-health crisis without any clear signal that the conversation has shifted registers. The system may respond poorly5,11 or well when one does.33

The rest of this Article focuses on Category 2. That is where the gap is widest: these systems invite the reliance people place in mental-health professionals, while the law imposes few corresponding duties on their developers. Category 1 already sits within a regulatory framework and at present almost no chatbots fall within it. Category 3 therefore raises issues that are more continuous with other harms associated with general-purpose conversational AI. Products liability law may develop in ways that address these risks without requiring the specialized duties we propose for Category 2, but the law remains too unsettled to know whether that will prove sufficient.

When it comes to Category 2, the status quo is hard to defend. Systems designed to provide mental-health support predictably receive disclosures of users in crisis yet carry almost no obligations except, in a few states, a duty to provide information directing users to crisis hotlines or other emergency support services. That this is the current state of the law seems less like a considered choice than a gap created

by the fact that these chatbots do not easily fit into categories established without these technologies in mind. Whatever duties are imposed should track how a system holds itself out. We describe a workable test below.

Drawing the Boundary

The test of whether a chatbot falls into Category 2 should depend primarily on how the developer holds out the chatbot and invites reliance, not how a particular user experiences it. The latter would be a very fact-specific that would be difficult to apply and would not give developers sufficient notice of the standards that apply to them.

This holding-out is inspired by, but does not replicate, the intended-use inquiry familiar from U.S. medical-device law. FDA regulation turns on whether a product is intended for a medical-device purpose, as evidenced by factors including labeling and promotional claims. Our inquiry asks a different question: whether, even without making claims sufficient to trigger device regulation, developers present a product in ways that invite users to rely on it for mental-health support.34 Products that hold themselves out in ways that deliberately invite the trust associated with professional mental-health services invite a different relationship from products marketed as writing assistants, coding tools, or search interfaces. The law should recognize that difference by assigning them different obligations.

Could a developer argue that, while the product holds itself out in this way, its marketing carefully includes disclaimer language disclaiming a therapeutic function? Such an approach would establish a regulatory safe harbor: developers who want to remain outside Category 2 would need to say so clearly and design their systems to match. We are skeptical that such a safe harbor is justified because it is unclear whether disclaimers actually work.

Merely including such language in product marketing should not suffice to qualify for such a safe harbor —the kind of boilerplate one frequently sees in drug advertising, for example. Instead, at the very least, one would need prominent disclaimers appearing frequently during use sessions, similar to California's requirement that operators frequently remind children that AI is not a human being. Even the more frequent in-session disclaimers may not actually make a difference for the way users experience the chatbot. Conversations with AI systems evolve. A user may begin by asking for help debugging code and gradually disclose loneliness, depression, or suicidal thoughts. A system that remembers personal information, expresses concern, validates emotions, and encourages continued disclosure may create expectations that conflict with a formal disclaimer. Character.AI opens every conversation with a red-lettered warning that "everything characters say is made up," and tells users to consult certified professionals; users still turn to its therapist-named bots.12 Whether well-designed frequent disclaimers can prevent this relational drift remains an open empirical question. Our own view is that disclaimer language should not engender a safe harbor unless and until the evidence decisively shows such language is effective. But we believe the question of regulatory design is closer. Even a regime that permits states to consider such safe harbors while imposing a background rule of duties on Category 2 chatbots would represent a major step forward in this area of law.

The Content of the Duty to Report for Category 2 Systems

A Category 2 system should be informed by duties informed by those imposed on clinicians facing comparable disclosures. But the analogy cannot be literal—many clinical duties depend on professional judgment that a chatbot does not possess.

For these systems, the natural reference point is the set of duties the law already places on the professionals they stand in for: the more an AI system substitutes for a human function, the more its regulation should resemble the rules governing the human who used to perform it.35,36 Concretely, lawmakers or courts could impose obligations on mental-health chatbots analogous to the mandatory-reporting statutes that bind clinicians—for suspected child or elder abuse, domestic-violence injuries, or gunshot wounds—and to a *Tarasoff-style* duty to warn an identifiable person when a credible, specific threat of violence emerges.

There are other problematic chatbot interactions, however, that go beyond the *Tarasoff* case. Some states have not extended *Tarassof* to cases involving self-harm rather than harm to another.37 But even when tort liability does not extend that far, clinical practice does some time impose affirmative obligations when a patient presents a serious risk of harm to self or others.38 The duties below draw on that broader structure. One asymmetry follows directly. A clinician ordinarily discharges obligations toward a suicidal patient from within the treatment relationship—by reassessing risk, adjusting the treatment plan, involving family with consent, or, in extreme cases, initiating hospitalization. A Category 2 chatbot has none of these capacities. It cannot treat, admit, or maintain continuity of care. Where a clinician's duty of care is satisfied internally, the functional equivalent for a chatbot must run outward. That means directing the user to a human being or service capable of intervening, since the chatbot itself cannot. This is not asking chatbots to do more than clinicians do; it is the same protective function, carried out through the only means available to a system that cannot deliver care directly.

Even so, these duties cannot be imported wholesale. Some depend on contextual clinical judgment—weighing a threat against everything known about a patient—that a chatbot cannot reliably make. For Category 2 systems, the law should therefore define concrete triggers for escalation rather than ask the system to make an open-ended judgment about whether a user is “dangerous.” Those triggers would include, for example, a stated suicide method and timeframe, a threat against a named or identifiable person, or a report of ongoing abuse of an identifiable child.

More specifically, we believe Category 2 AI systems should have five related duties:

- **Detection.** Developers should design Category 2 systems to identify defined indicators of acute risk (such as statements expressing suicidal intent, a method, or a timeframe), rather than respond only when users employ predetermined crisis phrases (such as “I want to kill myself”). This would require reasonable detection mechanisms, not a guarantee that the system will correctly predict which users will ultimately harm themselves or others.

- **Human review before external escalation.** When a system detects information that may meet a defined threshold for notifying a third party, emergency service, or public authority, an appropriately trained human should ordinarily review the flagged exchange and enough surrounding context to determine whether that threshold is met before identifiable user information is disclosed. Human review should be reserved for these high-risk cases rather than routine expressions of distress. Where waiting for review would itself materially increase the risk of imminent harm, lawmakers could permit immediate escalation subject to prompt human review afterward.
- **Proportionate crisis response.** Where the threatened harm is to the user, the response should be proportionate to the apparent risk. A system may direct the user to appropriate crisis or emergency-health services; when the threshold for external escalation is met, the developer should notify an actor capable of providing an effective crisis response. Law enforcement may sometimes be appropriate, but should not be the default where crisis or emergency-health services can respond effectively.
- **Independent audit.** Developers should open their detection and escalation performance to periodic outside review, and give independent researchers access to the data. Today developers hold all usage and outcome data, so whether these systems help, harm, or make no difference at scale is not merely unknown but unknowable to those outside the development team.38
- **Record-keeping.** Developers should preserve records sufficiently to determine how the system detected and responded to qualifying high-risk disclosures. Subject to appropriate privacy protections, developers should make those records available to public authorities responsible for enforcing statutory obligations and, where relevant, in discovery in litigation brought on behalf of an injured user.

Confidentiality is the real complication. For clinicians, a duty to report operates against a strong baseline of confidentiality grounded in professional ethics, privacy law, and other legal obligations.39 Reporting is therefore an exception to a broader rule of non-disclosure. HIPAA illustrates the structure. It does not itself impose a duty to report, but it expressly permits a covered provider to disclose protected health information to law enforcement or to others able to prevent harm where the provider believes in good faith that disclosure is necessary to prevent a serious and imminent threat, and separately permits notifying family or others involved in the patient's care. The obligation to act comes from state law and professional ethics; HIPAA supplies the permission that makes acting lawful. Category 2 chatbots generally lack an equivalent professional duty of confidentiality, even where consumer-privacy or health-data laws apply.38,40,41 That weakens the analogy to professional reporting: clinical reporting presumes confidentiality as the rule and disclosure as the exceptional response to defined risks. Importing the exception without the baseline would borrow only one side of that framework. A more coherent regime might therefore pair reporting duties for Category 2 systems with corresponding duties of confidentiality. Whether such a duty should attach, and in what form, is beyond the scope of this Article.

Enforcement

Lawmakers and courts could implement the duties proposed above through two principal mechanisms: statutory requirements imposed as a condition of market access, or public or private enforcement through litigation. FDA clearance is unlikely to supply the mechanism because Category 2 systems, by definition, generally fall outside medical-device regulation.34

Under the premarket clearance approach, by statute, compliance becomes a condition of entering the market. Failure to have the duties above in place would expose a developer to statutory damages or product withdrawal, whether or not anyone had yet been hurt. Because Category 2 systems generally fall outside federal medical-device regulation, federal device law should not ordinarily preempt state market-access requirements on that basis, although products that also contain FDA-regulated device functions could present additional preemption questions.

The adjudication approach requires someone to initiate a lawsuit. This could be a so-called "private right of action" -- brought by an injured party or their estate or guardian -- or it could be a "public right of action" -- brought by the attorney general of a state, for example. In existing state statutes on AI chatbots, while not applying their scheme to duties to report, we see examples of each. California's companion-chatbot law lets an injured user sue directly, for the greater of actual damages or $1,000 per violation.27 New York's equivalent gives users no such private right of action and leaves enforcement to the attorney general, at penalties of up to $15,000 a day.29 Lawmakers could plausibly extend either state's approach to duties to report as well. Existing negligence and product-liability law may provide another path: in Garcia v. Character Technologies, a federal court allowed negligence and design-defect claims involving a conversational AI system to proceed, by surviving the defendant's motion to dismiss, although the case later settled.42,43

Under either adjudication approach there would be a series of subsequent design choices: does failing to meet the duties to report listed above create a rebuttable or irrebuttable presumption that the developer fell below the standard of care? What level of awareness of the risks on the part of the developer should be required for what kind of liability?

We believe we have too little information at the present moment to be confident about the optimal design choices -- such as whether to favor the statutory preclearance or adjudication approaches and the best way to tailor either. Instead, we hope to see experimentation by different states, data gathering on effects of different enforcement schemes, and monitoring to learn what set of design choices are optimal. That experimentation should also examine whether reporting requirements deter candid disclosures, a risk that reinforces the case for narrow triggers, clear notice, and proportionate escalation.

Conclusion

AI chatbots now routinely receive disclosures that, if made to a clinician—or, in some jurisdictions, to another mandated reporter—could trigger affirmative legal duties to report. The law should recognize similar duties for chatbots under some circumstances. But the answer is not a single reporting rule for

every conversational AI system. Different chatbots perform different functions, create different expectations, and warrant different legal obligations.

This Article has proposed a functional framework for identifying when reporting duties should arise and explored the institutional choices involved in enforcing them. We argue that these questions should now be confronted directly. As conversational AI increasingly occupies roles once held by clinicians and other trusted intermediaries, continued legal silence is no longer good enough.

Declarations

Data availability

No datasets were generated or analysed during the current study.

Competing Interests

Prof. Cohen is a member of the Bayer Bioethics Council; a bioethics advisor for Bexorg; an advisor for World Class Health and Manhattan Neuroscience LLP; recently concluded service as the chair of the ethics advisory board for Illumina; was also compensated for speaking at events organized by Philips with the Washington Post, as well as the Doctors Company, attending the Transformational Therapeutics Leadership Forum organized by Galen Atlantica; and was retained as an expert in health privacy, gender-affirming care, and reproductive technology lawsuits. No other disclosures were reported.

Author Contribution

S.P., V.J. and I.G.C. conceived the article. S.P. and V.J. drafted the manuscript. I.G.C. contributed to the legal analysis and critically revised the manuscript. All authors read and approved the final manuscript and are accountable for its content.

Code availability

Not applicable.

Funding and all other required statements

S.P. and I.G.C. were supported in part by a Novo Nordisk Foundation grant for a scientifically independent International Collaborative Bioscience Innovation & Law (Inter-CeBIL) Programme (grant NNF23SA0087056).

References

1. Complaint at ¶¶ 50, 69–78, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. S.F. Cnty., filed 26 Aug. 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf.
2. Gottfried, J. et al. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact (Pew Research Center, 2026). https://www.pewresearch.org/wp-content/uploads/sites/20/2026/06/PI_2026.06.17_Americans-and-AI_REPORT.pdf.
3. De Freitas, J., Uguralp, A. K., Oğuz-Uğurlap, Z. & Puntoni, S. Chatbots and mental-health: insights into the safety of generative AI. J. Consum. Psychol. 34, 481–491 (2024). https://doi.org/10.1002/jcpy.1393.
4. De Freitas, J. & Cohen, I. G. Unregulated emotional risks of AI wellness apps. Nat. Mach. Intell. (2025). https://www.nature.com/articles/s42256-025-01051-5.
5. Cohen, I. G. & Palmieri, S. AI therapists vs companions: wellness, licensure, and liability. Am. J. Bioeth. 26, 56–58 (2026). https://doi.org/10.1080/15265161.2025.2608635.
6. Lucas, G. M., Gratch, J., King, A. & Morency, L.-P. It's only a computer: virtual humans increase willingness to disclose. Comput. Hum. Behav. 37, 94–100 (2014). https://doi.org/10.1016/j.chb.2014.04.043.
7. Croes, E. A. J., Antheunis, M. L., van der Lee, C. & de Wit, J. M. S. Digital confessions: the willingness to disclose intimate information to a chatbot and its impact on emotional well-being. Interact. Comput. 36, 279–292 (2024). https://doi.org/10.1093/iwc/iwae016.
8. O'Dowd, A. ChatGPT: more than a million users show signs of mental-health distress and mania each week, internal data suggest. BMJ 391, r2290 (2025). https://doi.org/10.1136/bmj.r2290.
9. Hui, V. et al. Perspectives and preferences of domestic violence survivors regarding digital platform and AI chatbot for help-seeking: a qualitative study. PLoS ONE 21, e0342453 (2026). https://doi.org/10.1371/journal.pone.0342453.
10. Xie, W. et al. Who will turn to chatbots for self-disclosure? Exploring conversation with chatbots, self-compassion, and existential meaninglessness in American and Hong Kong adolescents. AI Soc. 41, 4087–4100 (2026). https://doi.org/10.1007/s00146-025-02771-3.
11. De Freitas, J. & Cohen, I. G. Disclosure, humanizing, and contextual vulnerability of generative AI chatbots. NEJM AI 2, AIpc2400464 (2025). https://doi.org/10.1056/AIpc2400464.
12. Tidy, J. Character.ai: Young people turning to AI therapist bots. BBC News (4 Jan. 2024). https://www.bbc.com/news/technology-67872693.
13. Office of the Texas Attorney General. Attorney General Ken Paxton Investigates Meta and Character.AI for Misleading Children with Deceptive AI-Generated Mental Health Services (press release, 18 Aug. 2025).
14. Cal. Penal Code §§ 11160, 11161.9(a), 11166, 11172(a).
15. Child Welfare Information Gateway. Mandatory Reporting of Child Abuse and Neglect (U.S. Dep't of Health & Human Servs., 2023). https://www.childwelfare.gov/resources/mandatory-reporting-child-abuse-and-neglect/.

16. Gemayel, K. T., Hiro, M. E., Sullivan, R. R. & Payne, W. G. Review of statutory obligations for reporting ballistic injuries. Plast. Reconstr. Surg. Glob. Open 11, e4832 (2023).
17. Beggs v. State, Dep't of Soc. & Health Servs., 171 Wash. 2d 69 (2011).
18. State v. Strauch, 345 P.3d 317 (N.M. 2015).
19. Tarasoff v. Regents of the University of California, 17 Cal. 3d 425 (1976).
20. Slobogin, C. Tarasoff as a Duty to Treat: Insights From Criminal Law. 75 Univ. Cin. L. Rev. 645 (2006).
21. Johnson, R., Persad, G. & Sisti, D. The Tarasoff Rule: The Implications of Interstate Variation and Gaps in Professional Training. Focus 17, 435–442 (2019).
22. Thapar v. Zezulka, 994 S.W.2d 635 (Tex. 1999).
23. Johnson, R., Persad, G. & Sisti, D. The Tarasoff rule: the implications of interstate variation and gaps in professional training. J. Am. Acad. Psychiatry L. 42, 469–477 (2014).
24. Krikorian v. Barry, 196 Cal. App. 3d 1211 (1987).
25. Geraci v. Hamilton, 351 A.3d 24 (2026).
26. Scordato, M. R. Understanding the absence of a duty to reasonably rescue in American tort law. Tul. L. Rev. 82, 1447–1503 (2008).
27. Cal. Bus. & Prof. Code §§ 22601–22606.
28. S.B. 243, 2025–2026 Reg. Sess., ch. 677 (Cal. 2025).
29. N.Y. Gen. Bus. Law §§ 1700–1704.
30. Cohen, I. G. & De Freitas, J. Mitigating Suicide Risk for Minors Involving AI Chatbots—A First in the Nation Law. JAMA 335, 301 (2026). https://doi.org/10.1001/jama.2025.23744.
31. Srivastava, N. Liability for chatbots: a psychbot negligence case study and the need for reasonable human oversight. Torts Law Journal 28, 155–184 (2023).
32. Lee, H., Handler, R., Mungle, T. & Hernandez-Boussard, T. Building safer artificial intelligence mental health chatbots: a framework for transparency, evaluation, and shared accountability. J. Am. Med. Inform. Assoc. 33, 1538–1553 (2026). https://doi.org/10.1093/jamia/ocag078.
33. Pichowicz, W., Kotas, M. & Piotrowski, P. Performance of mental-health chatbot agents in detecting and managing suicidal ideation. Sci. Rep. 15, 31652 (2025). https://doi.org/10.1038/s41598-025-17242-4.
34. Simon, D. A., Shachar, C. & Cohen, I. G. Skating the line between general wellness products and regulated devices: strategies and implications. J. Law Biosci. 9, lsac015 (2022).
35. Haupt, C. E. Governing A.I.'s professional advice. McGill Law Journal 64, 665–682 (2019).
36. Balkin, J. M. The path of robotics law. Calif. Law Rev. Circuit 6, 45, 46, 55–60 (2015).
37. Lior, A. Tarasoff Meets the AI Age. Lawfare (20 Apr. 2026). https://www.lawfaremedia.org/article/tarasoff-meets-the-ai-age.
38. Bellah v. Greenson, 81 Cal. App. 3d 614, 146 Cal. Rptr. 535 (1978).

39. Ajunwa, I., Parikh, R. B. & Cohen, I. G. When patients share everything with an AI chatbot: risks and opportunities of large language models. JAMA 336, 279–280 (2026). https://doi.org/10.1001/jama.2026.9507.
40. 45 C.F.R. pts. 160, 164; 45 C.F.R. § 164.512(j)(1)(i); 45 C.F.R. § 164.510(b)(2); U.S. Dep't of Health & Human Servs., FAQ 2098, https://www.hhs.gov/hipaa/for-professionals/faq/2098/.
41. Marks, M. & Haupt, C. E. AI chatbots, health privacy, and challenges to HIPAA compliance. JAMA 330, 309–310 (2023).
42. Wash. Rev. Code ch. 19.373 (My Health My Data Act).
43. Garcia v. Character Technologies, Inc., 785 F. Supp. 3d 1157, 1180 (M.D. Fla. 2025).
44. Order, Garcia v. Character Technologies, Inc., No. 6:24-cv-01903-ACC-DCI, ECF No. 244 (M.D. Fla. Jan. 7, 2026).