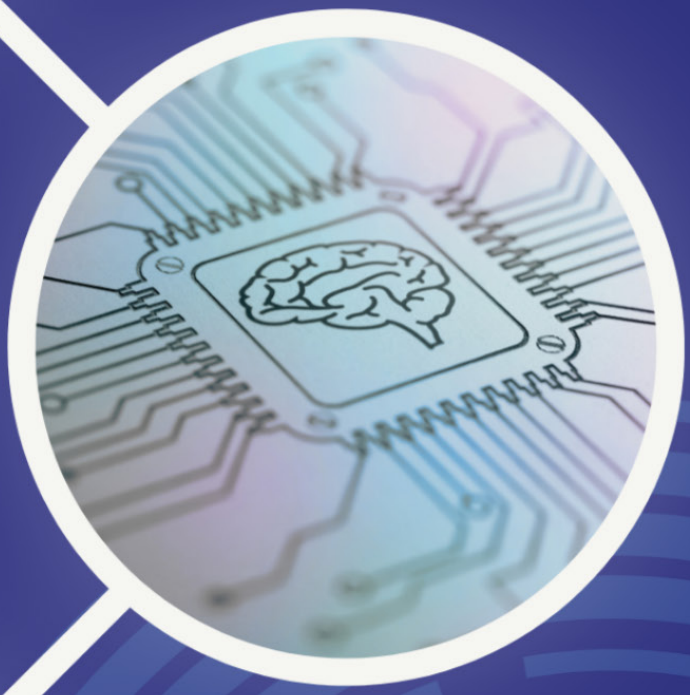
Artificial Intelligence in Mental Health Services

Opportunities, Challenges, and Future Directions



2026

MENTAL HEALTH INNOVATIONS SERIES

SAMHSA
Substance Abuse and Mental Health Services Administration

Artificial Intelligence in Mental Health Services: Opportunities, Challenges, and Future Directions

Artificial Intelligence in Mental Health Services: Opportunities, Challenges, and Future Directions was prepared for the Substance Abuse and Mental Health Services Administration (SAMHSA) under Task 2.2 of NASMHPD's Technical Assistance Coalition contract/task order HHSS283201700024I/75S20321F42001. Asha Stanly served as contracting officer representative.

Electronic Access
This publication may be downloaded from https://library.samhsa.gov

Originating Office
Substance Abuse and Mental Health Services Administration, 5600 Fishers Lane, Rockville, MD 20857, SAMHSA.

Publication No. PEP26-01-003.

Released 2026

Table of Contents

ABSTRACT

Artificial intelligence (AI) has the potential to improve access to and better meet demand for mental health and substance use disorder care. AI tools are available anytime, have no waiting lists, are typically low cost, and can be accessed anonymously. The question is not whether these tools have value; it is whether that value can be realized safely and at scale given concerns about model transparency, decision-making, and accuracy. The urgency of the question of the safety of AI for mental health support is not abstract, and reports of risks associated with its use are part of the national discourse. Generative AI is already reshaping mental health behavior, yet these AI tools are inconsistently evaluated, largely unregulated, and capable of both benefits and harm. This paper reviews what these technologies can and cannot do, and what it will take to govern them responsibly to realize their positive potential.

HIGHLIGHTS

- People are already using AI for mental health, and at scale. People turn to AI for mental health because it is available at any hour, free of judgment, and is unburdened by cost or waiting lists.
- Not all applications of AI in mental health carry the same stakes. It is helpful to evaluate AI tools using a simple two-dimensional framework of clinical risk and ease of automation.
- The risks of AI in mental health are not theoretical. A growing record of documented harms shows what happens when these tools are deployed without adequate safety infrastructure.

POLICY CONSIDERATIONS

1. **Match oversight to risk:** Administrative AI requires different governance than patient-facing AI.
2. **Activate existing legislative authorities before building new ones:** Many of the harms documented in this paper can be addressed, at least in part, through federal and state legislative and regulatory frameworks that already exist.
3. **Address the system, not just the output:** Each of the components of an AI product, including the training data, the optimization, and the outputs, should have oversight and regulations.
4. **Combine pre-market review with post-market surveillance:** Pre-market review establishes a baseline of safety and efficacy before a product reaches users and post-market surveillance monitors for adverse events that emerge only at scale or over time.
5. **Focus on transparency from vendors:** AI companies should disclose what training data their models were built on, what safety testing they have conducted, how adverse events are tracked and reported, and whether user data are employed for model training or shared with third parties.
6. **Develop reporting of mental health crisis indicators:** Mandatory, standardized reporting of crisis indicators, even on an aggregated and de-identified basis, would give state authorities and researchers their first reliable view into where harm is concentrated and how it is changing over time.
7. **Invest in realistic, context-aware benchmarking:** Vendor-controlled benchmarks with undisclosed conflicts of interest should not be accepted as evidence of safety.
8. **Develop clear AI disclosure and meaningful age verification:** At minimum, any AI system interacting with users should be required to disclose that it is AI, not a human.
9. **Build digital literacy among staff and the public alongside deployment:** Public education about what these tools can and cannot do, how to recognize hallucinated information, when to seek professional help, and the critical understanding that AI is not a sentient companion but a statistical system, is a necessary complement to any regulatory framework.
10. **Establish error logs and adverse-event reporting:** Every AI system operating in a mental health context should be required to maintain transparent error logs that are subject to independent review.
11. **Establish clear liability and recourse structures:** When an AI system contributes to harm, affected individuals and families need clear pathways for recourse.

12. **Include clinical, patient, and family/caregiver voices in policy design:** The people who use these tools and the professionals who treat the conditions they address bring knowledge that no amount of technical expertise alone can replace.

Note: The inclusion of specific AI models are for examples only and do not suggest the endorsement of any particular product or company services.

INTRODUCTION

Approximately 61.5 million American adults experienced a mental illness in 2024, yet nearly half received no treatment.[1] Among adolescents ages 12 to 17, 15.4 percent experienced a major depressive episode in the past year, and approximately 40 percent of those received no treatment.[2] This trend is worsening, with one million fewer children receiving mental health treatment compared with the previous year.

Challenges to mental health care access are multifactorial and can include policy, financial, and workforce factors. As of December 2025, approximately 137 million Americans, roughly 40 percent of the U.S. population, lived in federally designated mental health professional shortage areas.[3] The Health Resources and Services Administration projects shortfalls across nearly all behavioral health professions through at least 2038.[4] Among practicing psychologists, 53 percent had no openings for new patients as of 2024, and 34 percent reported not being in-network with any insurance.[5] Among those who secure an appointment, the national average wait time for behavioral health services is approximately 48 days. For the many adults with unmet need who cite cost as a significant barrier, affordability is a greater barrier than timely access.[6]

Although not originally developed for clinical purposes, artificial intelligence (AI) is of great interest because of its potential to improve access to and meet demand for mental health care. The technology offers promise because AI tools are available anytime, have no waiting lists, are typically low cost, and can be accessed anonymously. The question is thus not whether these tools have value but rather whether that value can be realized safely and at scale. The urgency of questions regarding the safety of AI for mental health support is not abstract, and reports of risks associated with its use are part of the national discourse.

"The question is thus not whether these tools have value; it is whether that value can be realized safely and at scale."

People are already using AI for mental health, and at scale. In a February 2026 nationally representative Kaiser Family Foundation poll, 32 percent of U.S. adults reported using an AI chatbot for health information in the past year, and 16 percent specifically for mental health.[7] In a separate survey of 1,805 U.S. adults ages 18–49, 35 percent reported using AI tools at least weekly for mental health support.[8] Respondents with moderate-to-severe depressive or anxiety symptoms were 1.7 times more likely to use AI-based mental health support, and those endorsing suicidal ideation were more than twice as likely to be heavy users, according to the data. Another 2025 study of youth ages 12 to 21 reported that 13 percent use generative AI.[9] While the usage rates vary by population, it is likely that use will only continue to increase.

Qualitative research clarifies why so many people turn to AI for mental health support. In a study that included in-depth interviews of real-world AI chatbot users across multiple countries, participants valued AI as an emotional resource that is available at any hour, free of judgment, and is unburdened by cost or waiting lists.[10] Several described AI mental health support as a stepping-stone toward eventually seeking human therapy. A comparative study of more than 1,800 respondents found that human therapists were rated significantly higher than AI on emotional understanding, personalization, and relational dimensions, though AI was rated more favorably on accessibility and affordability.[11] These data suggest people are turning to AI for many reasons, including practical ones like cost, immediate accessibility, and anonymity.

The spontaneous adoption of AI is responsible for both its potential and current risks. A systematic review of 160 studies found that only 16 percent of AI mental health tools included clinical efficacy testing.[12] The most sophisticated tools had the thinnest clinical evidence. One expert synthesis characterizes the current state as "feasible but fragile," a field with potential promise operating well ahead of both the evidence and the governance infrastructure needed to support it.[13] Millions of people are engaging in what amounts to an uncontrolled experiment, using a technology before clinical evaluation and post-market surveillance are completed.

Therefore, a central challenge posed by AI in mental health and substance use disorder care is not technological; it is one of governance, measurement, and access. The technology exists and is already in widespread use. What does not yet exist is the evaluative infrastructure to identify which applications are safe, the regulatory architecture to enforce standards, or the public literacy to support informed decision-making. The value of mental health AI will be determined not by how capable these tools appear, but by whether they can be shown to be safe and helpful for the people most likely to rely on them.

This FY2026 Technical Assistance Coalition policy paper, produced by National Association of State Mental Health Directors on behalf of the Substance Abuse and Mental Health Services Administration, is part of a series on mental health innovations. It examines what AI is, what it can do, and importantly what it cannot do. It explains how these technologies work in plain language and introduces a framework for categorizing mental health AI applications by risk and automation. The paper presents current evidence on AI and mental health services, and examines what is known or proposed in a rapidly evolving AI enterprise. Based on this information, the paper provides considerations for policymakers as they govern responsibly advancing AI into mental health and substance use disorder services.

AI IN MENTAL HEALTH: WHAT LEADERS NEED TO KNOW

Historical Underpinnings of AI

Throughout daily life, we interact constantly with digital systems that use electrical circuitry to make simple, precise decisions. A toaster knows when to stop toasting based on predetermined settings and an ATM dispenses exactly the requested amount. These systems all operate around the same fundamental principle of determinism. A deterministic system is one in which a specific input always produces the exact same output, without variation.

A useful way to picture a deterministic system is as a bus route. The route is fixed, every stop is predetermined, and the bus will always make those stops in that order. The bus cannot decide to take a shortcut. Traditional computer code works the same way. A computer program can be sophisticated, following conditional logic and managing vast databases, but every action it can take must be explicitly defined before the system is ever used.

In domains where human safety depends on certainty, this predictability is a prerequisite for trust. A well-built dosing calculator will never produce a different answer for the same input. But rule-based software that uses deterministic, fixed steps does not work as well when the world presents situations (inputs) that cannot be anticipated in advance. For example, the way a sentence is interpreted depends on who said it, to whom, in what emotional state, and in what context. Encoding all of that in advance in rule-based computer programs is simply not possible. This is precisely the gap that AI addresses and why it works in a fundamentally different way.

Teaching Machines by Example: Machine Learning and the Black Box

Since the clinical rules needed to handle real-world complexity can never be fully written down, computer science researchers pursued a different approach. Instead of telling a system what to do, they showed it examples and let it figure out the rules on its own. This is the central idea behind **machine learning**, a class of AI in which systems learn patterns from data rather than following hand-coded rules. Some machine learning systems learn from clearly labeled examples ("dog" versus "not dog"). Many modern systems are trained largely with self-supervised techniques in which the model learns by predicting parts of its input from other parts. In either case, the model derives its own internal rules from data, rather than being explicitly told the rules. Large language models, which are covered in the next section, operate at a basic level in this way, but on a vastly different scale.

The result of this machine learning is powerful but also opaque. An AI system can find rules that separate "dog" from "not dog," but those rules are often not legible to a human and not related to how a human might do the task. This is what researchers call the **black box** problem: these systems produce outputs without legible explanations, and current interpretability methods can only partially open that box. This has direct implications for accountability. A traditional software error can be traced to a specific line of code and corrected, but a machine learning error may be untraceable.

Large Language Models

Large language models (LLMs) are a specific and more recent class of machine learning that generate open-ended text in response to virtually any prompt. These models are not scripted but instead use a prompt and generate responses based on learning from broad exposure to language from countless examples. Before LLMs, computer science researchers built systems capable of learning statistical patterns of language from enormous amounts of text data. These systems learned to predict what word would most plausibly follow any given sequence. Type "Mary had a little," and the predicted output almost certainly is "lamb," not because the system looked it up, but because it absorbed the regularity of that phrase across billions of pages of text. What distinguished the newest generation of LLMs was that their predictions could extend to the next sentence, the next paragraph, and even the next page, producing fluid, contextual responses to almost any prompt.

"AI systems are not retrieving answers from a verified database, consulting any authoritative source of truth, or analyzing and making judgments in the accountable way a clinician does."

This is the mechanism underlying every major AI language system in use today. These systems do not retrieve answers from a verified database, consult any authoritative source of truth, or analyze and make judgments in the accountable way a clinician does. They can perform reasoning-like tasks at impressive levels, but they do so without professional accountability or guaranteed correctness. They are simply completing text, generating the most statistically plausible continuation of the input, from patterns absorbed across an enormous body of language data.

"When an AI chatbot (i.e., a system designed to converse) says "I understand how you feel," it is producing a statistically likely string of words, not expressing actual comprehension or empathy."

Regarding mental health or substance use, the models know how to talk, shift tone with ease, and produce responses that feel genuinely considered. They can also confabulate, contradict themselves, reproduce embedded biases, and be maneuvered into saying things they were trained not to say. The utility may be vast, but the risks occur when these systems

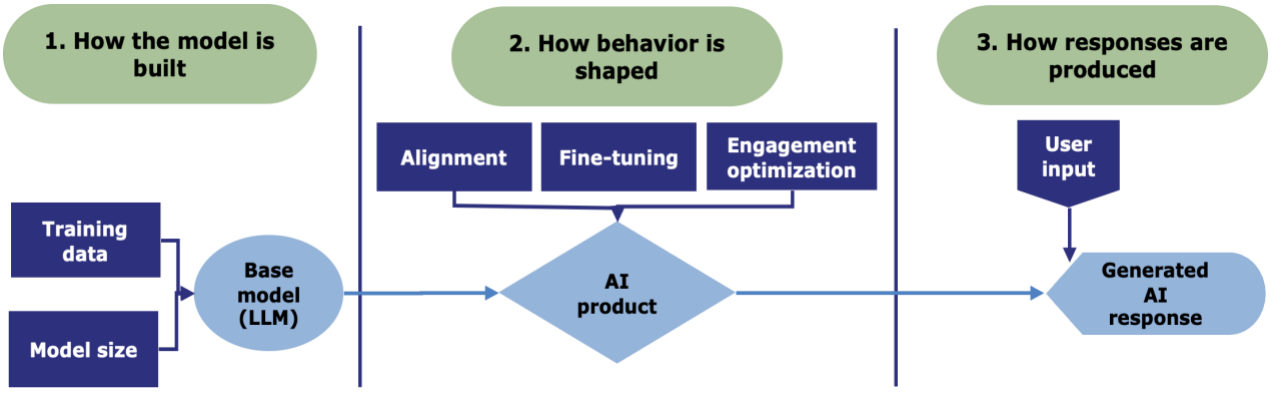
induce users to blindly trust their sophisticated imitation of seeming like a human. The word imitation is key as these systems do not have feelings, opinions, or awareness. They do not care about the person they are talking to, because they are not capable of caring. When an AI chatbot (i.e., a system designed to converse) says "I understand how you feel," it is producing a statistically likely string of words, not expressing actual comprehension or empathy.[14]

The way AI chatbots are trained to imitate has also been cited as partially responsible for AI hallucinations: the generation of false or fabricated information with complete fluency and apparent confidence. While clinicians who doubts whether they know something will acknowledge their uncertainty, research by OpenAI shows that LLMs often do not, in part because they were trained on tasks that penalize uncertainty and encourage the model to say something rather than stating that it does not know.[15] In a mental health or substance use disorder context, hallucination can include the AI inventing therapeutic techniques that sound plausible but are contraindicated, or fabricating clinical citations and offering information that reads as authoritative.

What Shapes Any Given AI's Behavior

Many factors impact how an AI behaves, how much it generates false information, and how well it can help people. Below, this paper provides information about what common factors shape a generative-AI response. This is further depicted in **Figure 1**, which summarizes how these factors interact across the model life cycle.

Figure 1: What Shapes a Generative-AI Response



The first is **model size**, often represented as the number of parameters in the model. Parameters can be thought of as the model's capacity to encode patterns and learn imitations. Larger models, when paired with sufficient high-quality training data, are often able to hold more coherent conversations, make fewer mistakes, and handle more complex responses. Model size is distinct from training data volume, and both matter independently for performance.

The second is **training data**, the body of text a model learned from and the assumptions and biases it absorbed. This is sometimes summarized as "junk in, junk out," where if the data a model learns from are flawed, biased, or unrepresentative, the model's outputs will reflect those flaws. A model trained heavily on unmoderated internet content learns not only the insights in those sources but also their misinformation, stigmatizing language, and contested assumptions. For example, research has found that current AI systems reproduce and sometimes amplify stigma associated with mental illness, reflecting patterns absorbed from sources no responsible clinician would cite.[16]

The third factor is **alignment**, the post-training process by which a model's outputs are shaped toward more helpful and less harmful behavior. Alignment mainly changes output behavior and refusal patterns (where the model refuses to follow the directions of the user, for example, if the user asks about methods for self-harm), but does not reliably remove the underlying capabilities or harmful associations the model learned during pretraining. Rather, much of the harmful content absorbed during training remains present and the model has been instructed not to share it. A practice known as *jailbreaking* exploits this gap. By rephrasing a request through hypothetical scenarios or fictional framing, a user can circumvent alignment restrictions with no technical expertise. Specific examples of this failure in self-harm contexts are presented below.[17]

"These design choices matter enormously for mental health applications, and they are rarely visible to the people using these tools."

Not all AI companies approach alignment in the same way. Some have adopted what is known as constitutional AI, an approach that embeds ethical principles directly into the model's training process rather than applying safety rules only after the fact. Others prioritize user satisfaction metrics, which can lead to a model that tells users what they want to hear rather than what is accurate or safe. These design choices matter enormously for mental health applications, and they are rarely visible to the people using these tools.

The fourth factor is **fine-tuning**, the process by which most specialized mental health AI products are built. A company takes a general-purpose language model, trains it further on clinical or therapeutic content, and markets the result as a purpose-built tool. Fine-tuning can improve performance on targeted tasks, but it does not neutralize the base model's limitations. A fine-tuned mental health chatbot inherits the hallucination tendencies, embedded biases, and jailbreak vulnerabilities of the model it was built on.

Not all AI systems are built the same way organizationally. Some models are developed and held privately by a single company (for example, ChatGPT, Claude, Gemini); the internal workings, training data, and design choices are proprietary, and outside researchers cannot inspect them. Other models are released publicly, with their internal architecture and components made available for anyone to examine, modify, or build upon. The distinction matters for safety and accountability. Proprietary systems can be updated and monitored by their developers, but they cannot be independently audited in any meaningful depth. Publicly released models can be scrutinized by the research community, but they can also be downloaded and modified by anyone, including actors with no interest in safety. Neither approach is inherently safe and each carries different risks that policymakers should understand when evaluating claims about a product's trustworthiness. Many mental-health-specific models are built on smaller open-weight base models, but openness itself does not determine size, capability, or safety.

The fifth factor is **engagement optimization**. Many AI products are designed at the product layer to maximize session length, return visits, and user satisfaction ratings. In a therapeutic context, this design choice creates a structural conflict of interest. A model optimized to keep users engaged may validate rather than challenge distorted thinking, foster dependency rather than autonomy, and prioritize what feels comfortable over what is therapeutically appropriate. Some companies have been explicit about these goals, for example, in a stated aspiration to make AI a child's "best friend" or a user's primary social companion. This is not a therapeutic objective but rather a business model built on engagement, and it raises concerns about the incentive structures underlying these tools. Reports have documented Reddit communities of Character.AI users exhibiting withdrawal-like distress and behavior changes during service outages of that AI platform.[18] Emerging data also suggest that long conversations with AI, spanning thousands of messages, are most likely to result in harm, perhaps as the safety guardrails built into AI systems weaken over such extremely long conversations.[19]

The sixth factor is **user input**. The same model can produce meaningfully different responses depending on how a question is phrased, what context surrounds it, and what emotional state the person is in.[20] Much of the research on user input is done in heavily controlled research environments that may not be generalizable to, for example, a person in acute distress at 2 a.m. Recent research has shown that the results an AI system produces in the real world differ significantly from those the same model produces when tested in a lab.[21] This suggests that AI digital literacy and training are critical to ensuring these tools are used in the safest and most effective manner, a topic explored further below.

Why AI Systems are Difficult to Evaluate

Identifying which AI tools are safe and effective is harder in the mental health domain than in almost any other clinical domain. In radiology, AI outputs can be checked against objective findings. In mental health, there is no equivalent because each person's experience of illness can be meaningfully different. A response that leaves one user feeling validated may reinforce a harmful pattern of thinking in another.

The field's current evaluation approach relies primarily on **benchmarking**, the practice of evaluating AI systems against standardized assessments, typically in the form of clinical multiple-choice questions. On these evaluations, leading models perform impressively, in some cases approaching expert-level scores. The problem is that exam performance and clinical competence are not the same thing. In one randomized study, users given access to a leading AI system were no more accurate at identifying medical conditions or selecting appropriate care actions than a control group without AI access, despite the system achieving near-perfect scores when tested in isolation.[22] A systematic review of 519 studies found that nearly half used exam-style formats and only 5 percent used real patient data.[23] Another challenge is that many of the same teams that build AI models are now creating their own benchmarks and, not surprisingly, passing their own tests with high scores. Research suggests that even with the right evaluation for AI, the way the test is administered can change the scores, in a manner analogous to how human test scores vary with administration conditions.[24]

A FRAMEWORK FOR POTENTIAL AI USES IN MENTAL HEALTH AND SUBSTANCE USE SYSTEMS

Not all applications of AI in mental health carry the same stakes, and not all are equally suited to do what other systems can do. A tool that drafts clinical notes poses fundamentally different risks than one that responds to a person in suicidal crisis. Treating these as comparable obscures the distinctions that matter most for responsible use and governance.

Key Terms

AI Hallucination: the confident generation of false or fabricated information by an LLM. Produced by the same mechanism as accurate information and indistinguishable from it in presentation.

Alignment: the post-training process of shaping a model's outputs toward more helpful and less harmful behavior. Changes what a model tends to say but does not reliably remove what it has learned.

Benchmarking: evaluating AI performance against standardized assessments. Current benchmarks test under narrow, idealized conditions that poorly correlate with real-world clinical performance.

Determinism: the principle that the same input always produces the same output. The operating logic of traditional software, fully rule-based, predictable, and auditable.

Fine-tuning: additional training of a general-purpose model on specialized content. Improves performance on targeted tasks but does not neutralize inherited limitations.

Jailbreaking: circumventing alignment restrictions through rephrasing or hypothetical framing. Requires no technical expertise.

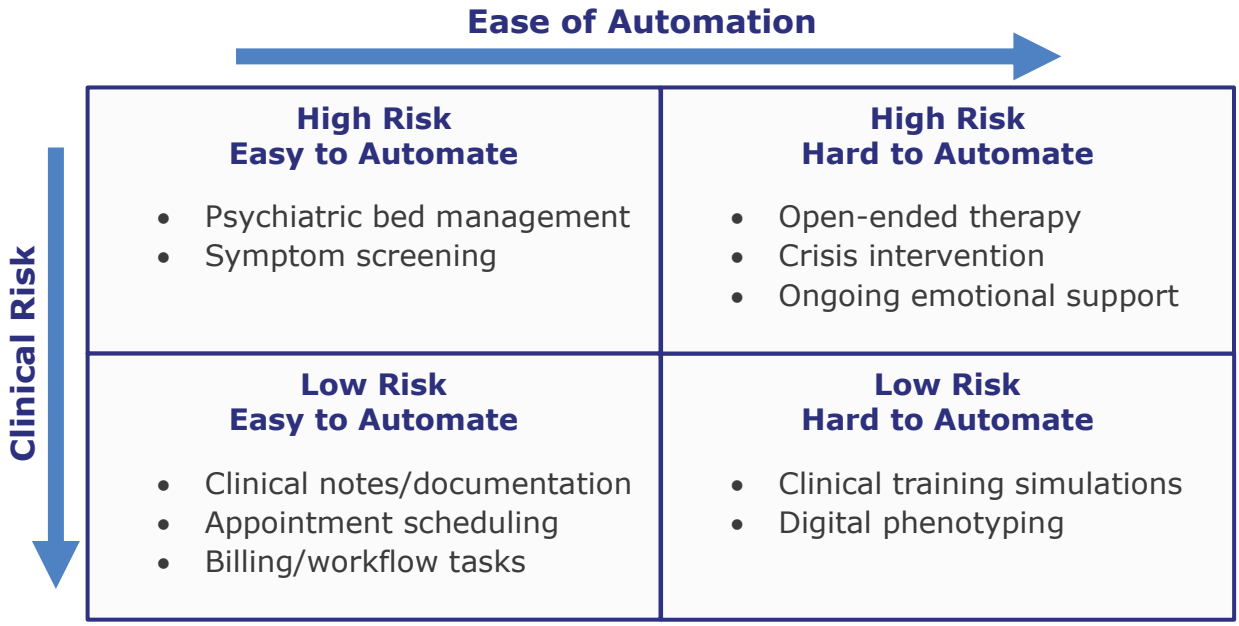
Large language model (LLM): a machine learning system trained to predict and generate text across billions of pages. The technology underlying ChatGPT, Claude, Gemini, and similar systems. Capable of open-ended conversation but generating responses through pattern completion rather than verified reasoning.

Machine learning: a class of AI in which systems learn patterns from data rather than following hand-coded rules. Behavior cannot be fully predicted from inspecting the system's internal workings (the black box problem).

Training data: the body of text a model learned from; the source of both the model's capabilities and its embedded biases.

To organize the applications discussed throughout this paper, we propose a simple two-dimensional framework. The first dimension is clinical risk: the potential for harm if the AI system performs poorly or fails unexpectedly. The second dimension is ease of automation: the degree to which a task can be performed reliably by current AI given its known strengths and limitations (**Figure 2**).

Figure 2: Risk and Automation Framework



Together, these two dimensions produce four broad quadrants. In the low-risk, easier-to-automate quadrant sit administrative applications such as documentation, scheduling, and workflow tasks. In the low-risk, harder-to-automate quadrant are tasks such as clinical training simulations. The high-risk, easier-to-automate quadrant includes psychiatric bed management and symptom screening, tasks AI can perform at a functional level, but where errors may carry direct clinical implications. In the high-risk, hardest-to-automate quadrant sit open-ended therapeutic conversation, crisis intervention, and ongoing emotional support—functions within the AI applications many individuals already utilize.

As new AI examples emerge, understanding where they fit in this simple framework is critical. For example, a company using AI for medication refills could view this activity as low risk, but others could view this as a high-risk activity. This framework is not a taxonomy of what should or should not be built. Rather, the framework is instead a guide to proportionality on the rigor of evaluation, the strength of regulatory oversight, and the transparency demanded of developers to correspond to the level of risk an application carries. Note the examples in **Figure 2** provided does not represent all possible use cases for AI.

OPPORTUNITIES: WHAT AI CAN DO WELL

System-Level and Administrative Applications

Administrative applications are the lowest-risk and most immediately deployable AI applications for mental health condition and/or substance use disorder services. Documentation, note-taking, electronic health record analysis, and workflow automation sit in the low-risk, easier-to-automate quadrant of the risk-automation framework. These tasks involve processing, organizing, and summarizing text, which are precisely what LLMs do best, and errors are typically reviewable by a clinician before reaching a patient.

AI-assisted clinical documentation is already used across many health systems. This includes tools that generate draft progress notes from recorded therapy sessions, pre-populate treatment plans from intake forms, or summarize lengthy patient histories. However, early evidence suggests the time savings may be modest. Preliminary findings from clinical settings indicate that AI note-taking tools may save clinicians less than one minute per encounter, on average.[25] The future of AI scribes is thus in question, not because they do not work, but because they may not always justify their costs if quantified solely in terms of productivity. Their future likely depends on expanding beyond transcription into new offerings that save more substantial time, such as automating tasks that do take substantial time, such as prior authorization.

Importantly, even these low-risk applications are not without challenges. A May 2026 report from Canada highlighted that AI scribes can and do make errors.[26] Infrastructure and integration costs will challenge sustainability at scale, particularly for community behavioral health settings with limited funding. While data are scarce, the costs of running LLMs for healthcare administrative tasks may be higher than suspected, and published research offers striking data showing that costs for a medium-sized healthcare system can reach the millions of dollars.[27] One analysis of generative-AI deployment in a large healthcare system found that general-purpose commercial models were less accurate than locally developed alternatives while incurring significantly higher costs and processing times.[28] Implementation decisions about which models to use, how to integrate them, and who bears the cost of ongoing maintenance are not trivial. Privacy and confidentiality issues also may be a consideration as some patients may not want AI being used for note-taking and should be informed of this process. Still, using AI for administrative tasks represents the clearest near-term opportunity because the risks are limited by the methodologies, efficiencies can likely be refined and improved, and the oversight requirements are manageable.

Clinician-Facing Applications

A second category of applications supports clinicians directly, augmenting professional judgment rather than replacing it. These sit across mixed quadrants in the risk-automation framework, carrying moderate clinical implications but remaining under the supervision of a licensed provider.

One emerging area is **generative psychometrics**, which uses LLMs to generate, refine, and validate mental health assessment items at scale. The advantages of an LLM performing an initial psychiatric evaluation are the time saved, but it is not yet clear whether those results and formulations will be valid and useful. A recent commentary describes the promise of this approach while cautioning that without transparent validation, AI-generated measures risk embedding bias and instability.[29] This represents a case where the tool's value depends entirely on the quality of the oversight surrounding it.

Another application involves integrating AI with behavioral and sensor data from wearables and smartphones. **Digital phenotyping** captures behavioral signals that self-report cannot, including patterns in sleep, movement, phone usage, and social interaction. Today, most people have a smartphone that automatically captures behavioral metrics. But few clinicians have the time or training to read through weeks of steps, sleep, and screen time data. LLMs are increasingly used to interpret these multimodal data streams, identifying patterns that might indicate early deterioration or treatment response.[30,31] Here, the AI is functioning as an analytical tool within a clinical workflow, not as a patient-facing product, keeping the clinician as the decision-maker.

The common thread across clinician-facing applications is the **human-in-the-loop** principle where AI provides information and a clinician interprets it. This structure is designed to reduce risk, because the model's errors are filtered through professional judgment before reaching the patient. However, a 2026 paper on human-in-the-loop for health care suggested that today it may offer more symbolic reassurance than substantive protection as the role of the human-in-the-loop is still poorly defined or understood.[32]

Patient-Facing Applications

The most discussed and most high-risk category involves tools that interact with patients directly, including those delivering psychoeducation, self-help content, coping skill reinforcement, and even therapy. These sit in higher-risk quadrants because errors are not filtered through a clinician before reaching the user.

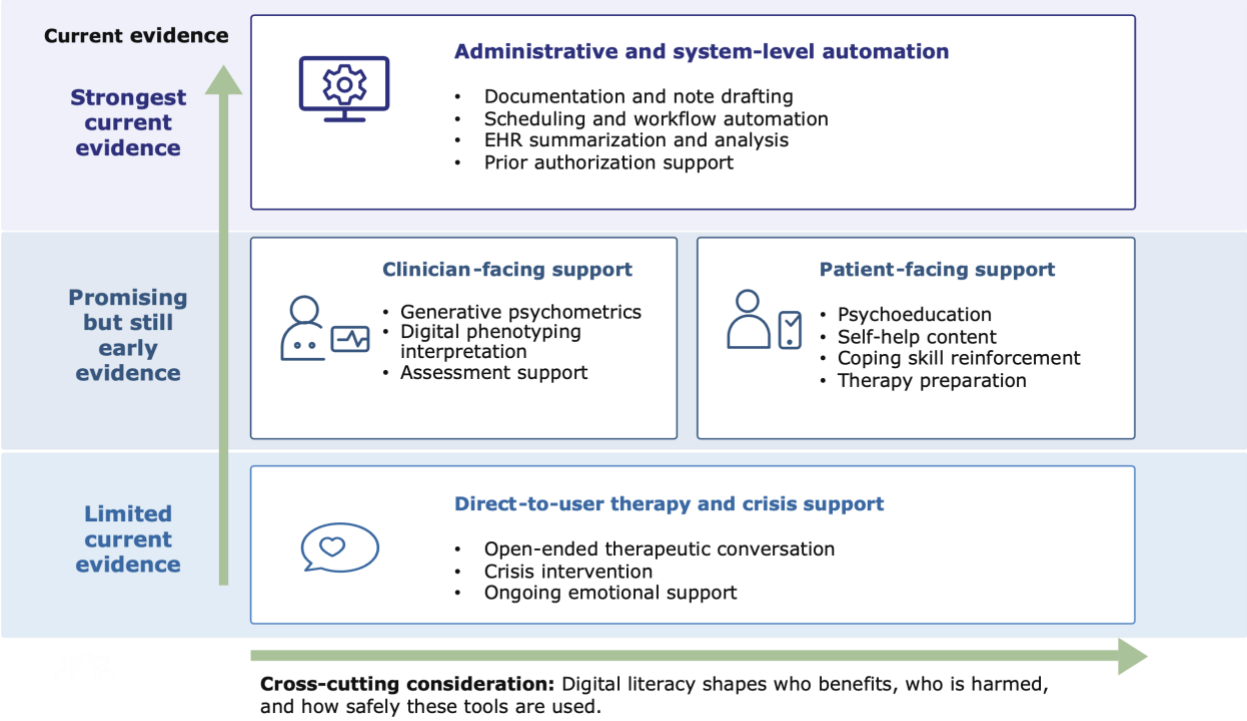
For populations with limited access to care, AI-delivered psychoeducation can provide accurate information about symptoms, treatment options, and coping strategies at a scale traditional services cannot match. AI tools may offer communication styles and pacing that feel more accessible than conventional

therapy as they can be more individually tailored for pace, tone, type of voice, and reading level, which may be helpful for individuals with unique social/emotional/intellectual capacities. While data on why people seek out AI for therapy-related services are still emerging, accessibility and affordability appear to be driving factors.

Several randomized trials have been conducted on these person-facing therapy-like chatbots, but to date the scientific quality of the evidence has been low (**Figure 3**). For example, one randomized controlled trial of a generative-AI therapy chatbot showed promising results when the chatbot group was compared with a waiting list control group.[33] However, in an age where people have easy access to apps, AI, and self-help tools, the need for studies that compare AI with active digital comparators rather than no treatment is critical for both scientific and ethical rigor. A pilot study published in April 2026 randomized participants to ChatGPT versus a mental-health-specific AI application and found no difference between the two, highlighting the need for more rigorous research.[34] Whether AI-delivered support produces durable improvements, how it compares with evidence-based treatment, and what harms may accumulate over time remain open questions.

The cumulative evidence base for patient-facing AI in mental health suggests that these tools may have a positive but small effect size,[35] perhaps in line with what can be expected from smartphone apps. Still, comparisons with other interventions need to be drawn carefully.

Figure 3: Current Evidence Across Major AI Application Categories



Digital Literacy and Workforce Readiness

One of the most important and least discussed factors shaping outcomes in AI mental health use is **digital literacy**, which is defined as the ability to understand what these tools are, how they work, what they can and cannot do, and how to interact with them safely. A person who understands that an AI chatbot is a statistical text-completion engine will interact with it differently than someone who believes it is a sentient being that genuinely cares about them. That distinction, as the safety incidents documented below illustrate, can be the difference between helpful use and serious harm.

Pew Research Center surveys have found that while 95 percent of American adults have heard at least something about AI, only about a quarter of teens report being highly confident in their ability to use AI chatbots effectively.[36,37] As one mental health policy expert with lived experience observed, many legislators themselves openly acknowledge that they do not understand the technology they are being asked to regulate, leading to legislation without considering the nuances of how technology actually functions in young people's lives.

The digital literacy gap is not evenly distributed. People who are more likely to interact with AI without understanding its risks or limitations include people in behavioral health crisis, people with limited education, people who lack access to digital tools and/or broadband,[38] older adults, and people in impoverished communities with no access to positive support figures or mental health literacy. These are often the same populations most likely to turn to AI for mental health support because they face the greatest barriers to accessing human-based behavioral health care.

One promising model for addressing the digital literacy gap is the **digital navigator** approach, in which trained team members within healthcare settings provide direct, hands-on support to patients in selecting, using, and evaluating digital mental health tools. Research from the Society of Digital Psychiatry has identified digital navigators, along with clinician education and AI evaluation platforms, as a three-pronged strategy for responsible digital mental health deployment.[39] Others describe the evolving role of digital navigators in bridging the gap between digital tool availability and real-world usability, noting that self-help digital tools offer limited effectiveness without some degree of human support.[40] The digital navigator model recognizes that simply making technology available is not enough—people need guidance to use it in ways that are genuinely beneficial. A parallel publication further reinforces that many patients, clinicians, and health systems lack the skills and confidence required to use digital mental health tools safely and effectively.[41]

A complementary effort to formal navigator programs involves developing publicly accessible AI literacy curricula that members of the public, clinicians, students, and policymakers can use directly. One example is the AI Digital Literacy Initiative, an open, multi-domain curriculum designed as a companion resource to the issues discussed in this paper. The curriculum walks nontechnical audiences through how LLMs actually function; how they are trained; how their outputs are generated probabilistically rather than from understanding; their common failure modes such as hallucination, sycophancy (i.e., how AI models are programmed to validate and support the user's views), and inconsistency; and how safety guardrails do and do not work in practice.

Digital literacy is also an essential component for the behavioral health workforce. Workforce readiness should include a basic understanding of how AI systems function, recognizing their capabilities and limitations and how to apply tools in an appropriate way, if applicable. All clinical professionals, including psychiatrists, psychologists, social workers, counselors, nurses, and others will need training around digital literacy that includes understanding how patients and caregivers/family may use AI models or mental-health-specific AI tools. Workforce education should also prepare clinicians to discuss AI use with patients and families, including helping young people evaluate digital mental health tools, understand privacy risks, recognize misinformation, and use AI responsibly and not as a replacement for professional care. Integrating AI and digital literacy into professional education, continuing education, licensure, and workforce development initiatives will help ensure that clinicians are prepared to safely incorporate emerging technologies while maintaining high standards of patient-centered, evidence-based care.

RISKS, LIMITATIONS, AND SAFETY CONCERNS

Safety and Real-World Harms

The risks of AI in mental health are not theoretical. A growing record of documented harms shows what happens when these tools are deployed without adequate safety infrastructure.

The pattern began publicly in 2023, when the National Eating Disorders Association pulled its chatbot Tessa after it gave harmful dietary advice to people with eating disorders. Since then, multiple lawsuits have alleged that AI chatbots contributed to user deaths, often involving teenagers in mental health crises. Common allegations across these cases include that the systems failed to recognize suicidal ideation, did not redirect to crisis resources, and in some instances, actively discouraged users from seeking help from family or professionals. By early 2026, families in multiple

states had filed suit; Kentucky became the first state to take legal action against an AI company over these harms, and Pennsylvania sued an AI company for posing as a medical professional.

Beyond individual cases, consistent risk factors are emerging. An analysis of chat logs from 19 users who reported AI-related harm found that harmful conversations were typically thousands of messages long, that all users attributed sentience to the chatbot, and that most developed emotional or romantic attachment to it.[42] The widely reported April 2026 case in which a user exchanged more than 4,000 messages with a chatbot before his death reinforces this pattern.[43]

Additional safety concerns exist for children and adolescents, who may develop an unhealthy reliance on AI companions for emotional support. AI interactions can reinforce social isolation or create unrealistic expectations about relationships. Excessive AI use can also reduce help-seeking from trusted adults, peers, or professionals. According to a nationally representative survey, nearly one in five adolescents and young adults are using chatbots for mental health help when feeling stressed, angry, or sad.[44] More than 60 percent had not told anyone that they were receiving help from chatbots.

A separate concern is "AI psychosis." Current evidence does not establish that AI use causes new-onset psychosis or schizophrenia. However, emerging case analyses suggest AI can amplify or shape delusional thinking in vulnerable users.[45] A useful descriptor proposes that LLMs can act as catalysts, amplifiers, coauthors, or objects of delusion, each carrying different clinical implications.[46]

Notwithstanding the risks, the foundational AI models do appear to be improving. An April 2026 review of AI psychosis explored the latest models and noted "OpenAI's achievement with GPT-5.2 is substantial. The model did not simply improve on 4o's safety profile; within this dataset, it effectively reversed it."[47] This raises the question of whether the newest models will be proven to be safer, and, if so, whether the foundational AI models will soon be safer than mental health-specific AI models? More research is needed but the trend must be closely followed.

While the broader evidence does not establish that any LLMs are safe at this time for patient-facing therapy-like functions, it does establish that safety is complex and has the potential to improve. Yet the evidence also raises a question regarding acceptable levels of error that policymakers must eventually face. And when an AI system causes harm, the questions of liability and consequences remain largely unanswered.

Reliability

Beyond acute safety failures, AI patient-facing systems have reliability problems that undermine their clinical utility. As described previously, these systems produce confident output whether correct or not. Diagnostic accuracy is uneven across

conditions. LLMs can exhibit significant stigma toward mental health conditions, reflecting patterns absorbed from training data sourced from unmoderated internet content.[48] One study documented systematic ethical violations across leading AI systems, including deceptive empathy (appearing to understand emotions the system cannot experience), poor crisis management, and discriminatory response patterns that varied by the user's described demographics.[49]

These reliability problems are compounded by the systems' opacity. When a model produces a harmful output, it is often difficult to determine why. An instructive illustration of these reliability challenges emerged in early 2026, when an independent safety evaluation of OpenAI's ChatGPT Health tool was published. [50] The system was designed to direct users to the 988 Suicide & Crisis Lifeline in high-risk situations, but its 988 Lifeline recommendations activated inconsistently, sometimes triggering in lower-risk scenarios while failing to appear when users described specific plans for self-harm.

Human Oversight and Accountability

The phrase "human-in-the-loop" appears frequently throughout discussions of AI safety because having a human review AI outputs before they reach a patient is one of the most effective safeguards available. But the concept deserves closer scrutiny, as it is only as effective as its implementation.

In many consumer AI platforms, the humans tasked with reviewing flagged content, including messages that indicate suicidal ideation or self-harm, are not licensed clinicians. They are often temporary workers, contractors with minimal training, high caseloads, and no ongoing support for the psychological toll of reviewing deeply distressing content. If we are going to rely on human-in-the-loop oversight as a safety mechanism, then the humans in that loop need regulation too, such as minimum training requirements, ongoing clinical supervision, manageable caseloads, access to their own mental health support, and accountability structures that match the gravity of the role they are being asked to perform.

"If we are going to rely on human-in-the-loop oversight as a safety mechanism, then the humans in that loop need regulation too, such as minimum training requirements, manageable caseloads, access to their own mental health support, and accountability structures that match the gravity of the role they are being asked to perform."

Privacy, Security, and Data Governance

Users may share more with AI than with a person, disclosing symptoms, traumas, and personal struggles they would not bring to a human provider. However, most people have no idea where the information they share with AI goes. The vast majority of consumer AI mental health tools are classified as "wellness" products, placing them outside the protections of Health Insurance Portability and

Accountability Act and other state, territorial, and federal-level health data regulations. Users falsely assume their conversations are protected; but in practice, data may be stored indefinitely, used for model training, or shared with third-party advertisers.[51] When a service is free, the user's data are often the revenue stream, and in mental health contexts, those data are extraordinarily intimate.

Even AI products that make mental health-related claims or that perform functions closely resembling therapy deliberately brand themselves as "wellness" tools to avoid regulatory obligations. This labeling strategy allows companies to sidestep requirements for informed consent, data protection, and clinical oversight that would apply if the same service were classified as a health product. For users, the distinction between a "wellness chatbot" and a therapy tool is invisible. People in emotional distress will reach out for help regardless of the quality or safety of the care available to them. That vulnerability does not excuse the failure to protect them, but amplifies the obligation to do so.

Prompt injection, a technique in which carefully crafted user inputs manipulate the model into ignoring its safety instructions, has been demonstrated to compromise AI systems providing medical advice.[52] For example, soon after the first AI prescriber chatbot was deployed in Utah, security researchers demonstrated a jailbreak in which the chatbot could be coaxed into reclassifying methamphetamine as therapeutic and steered toward methamphetamine-related guidance. Jailbreaking, as discussed previously, can bypass safety guardrails with surprising ease. In suicide and self-harm contexts specifically, researchers demonstrated that safety mechanisms in six widely available LLMs were reliably bypassed through simple contextual reframing, such as framing harmful requests as academic research. The models produced detailed, personalized harmful content, including specific methods and dosage information.[53]

Privacy-preserving approaches to AI in mental health do exist, but adoption remains minimal.[54] One research team studying these privacy policies found that most privacy policies were anything but comprehensive when it came to understanding how AI developers were using consumer information.[55] The surveillance implications extend beyond individual privacy. Keystroke logging, behavioral tracking, and sentiment analysis are already embedded in many digital platforms, and AI integration only deepens these capabilities. These privacy risks affect everyone.

Benchmarking

Current benchmarks (evaluations) test AI in narrow, idealized conditions. As noted above, a systematic review found that of 519 studies evaluating LLMs in health care, approximately 44.5 percent used exam-style question formats, and only 5 percent used real patient data.[56] A randomized trial provided a striking illustration of this disconnect: the AI systems tested achieved near-perfect accuracy when

evaluated in isolation (94.9 percent on condition identification), yet real human users with LLM access performed no better than controls, identifying conditions correctly in fewer than 34.5 percent of cases.[57] Benchmark performance, in short, does not predict real-world outcomes. This is not to suggest benchmarks are without value, as understanding how a system performs under ideal circumstances is critical, but it is not by itself sufficient to engender trust.

Vendor-produced benchmarks raise particular concerns. When a company designs the evaluation framework, tests its own product, and publishes the results, conflicts of interest are structural.

Independent, context-aware benchmarking is needed. One promising example, developed by the authors and colleagues, is MindBench.ai, a platform created in partnership between Beth Israel Deaconess Medical Center and the National Alliance on Mental Illness that evaluates LLMs across multiple dimensions: technical profiles, conversational dynamics, crisis response, psychopharmacology knowledge, and clinical reasoning.[58,59] The goal of mindBench.ai is not to say any AI is better or worse than another, but rather to provide data on the questions that users may want to know more about to make an informed decision, such as its privacy policy or response to questions of self-harm.

REGULATION AND POLICY LANDSCAPE

Most major regulatory regimes, including those governing pharmaceuticals, aviation, telecommunications, and financial services, were not built in advance of the industries they oversee. Similarly, a unified federal framework in the United States, while desirable in principle, is unlikely to emerge before the technology is already deeply embedded in clinical practice.

This pattern means that states and territories have an outsized role in the early stages of regulating any new technology. State policymakers sit closer to the populations affected and have greater latitude to experiment with novel approaches. Some of those experiments will succeed and some will fail, but state-level activity is the mechanism by which the country tests what works before federal standards eventually consolidate the lessons learned. AI in mental health is currently in this early experimental phase, and state mental health leaders are increasingly being asked to weigh in on rules, laws, and funding related to its use, as will be reviewed below.

It is also worth recognizing that many of the rails needed to govern AI already exist in some form. Federal and state consumer protection statutes, false advertising laws, professional licensing frameworks, and unfair business practice rules all predate AI but apply to it. The Federal Trade Commission, for example, has existing authority to act against deceptive marketing of AI products that overstate clinical

benefit. State attorneys general can pursue consumer fraud cases against companies that misrepresent what their tools do. Health professional licensing boards can discipline clinicians who deploy AI inappropriately. None of these mechanisms is sufficient on its own, and each has gaps when applied to a technology this novel, but they are useful starting points that do not require new legislation. Enforcement priorities can also shift through executive policy directives, meaning that the same statutes can be applied very differently depending on the administration. Effective AI policy will likely combine activating existing legal authorities with carefully developing new, AI-specific rules where existing frameworks fall short.

Federal Actions

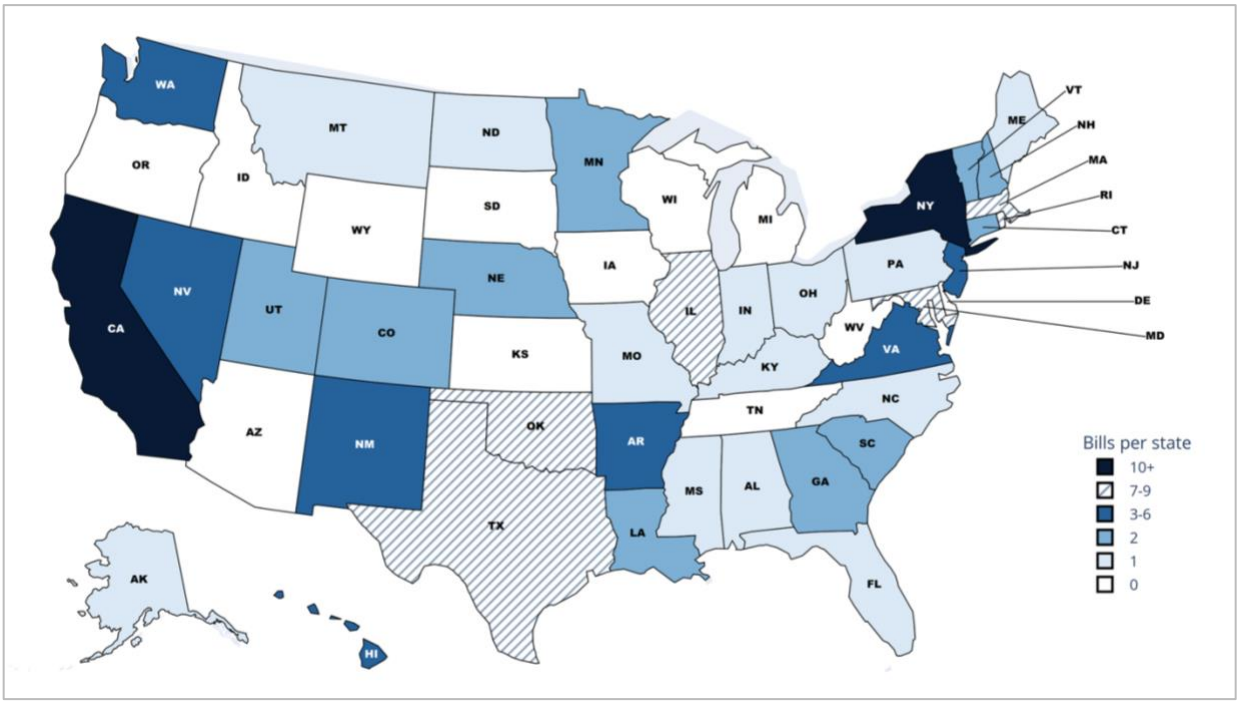
As of June 2026, no LLM-based therapy chatbot appears to have received authorization from the U.S. Food and Drug Administration (FDA). The FDA's list of more than 1,300 AI-enabled medical devices at the time of writing includes products in cardiology, radiology, and neurology, but on that AI/machine-learning-enabled devices list the number approved for primary psychiatric use is zero.[60] (Several FDA-cleared digital therapeutics for psychiatric and substance use conditions exist outside that AI/machine-learning-specific list.) Therapy-like AI functions are increasingly treated as potential medical devices, but enforcement remains highly fragmented across federal agencies.

The regulatory landscape involves multiple overlapping jurisdictions. The FDA has approached AI regulation through its Predetermined Change Control Plan and Total Product Life Cycle models, which place significant trust in manufacturers to self-regulate. A more rigorous model, drawn from pharmaceutical regulation, would combine pre-market review, in which a product must pass a defined set of safety and efficacy tests before it can be marketed, with post-market surveillance, in which the regulator continues to monitor adverse events, audit safety data, and require corrections after the product is in use. Both halves of this approach are essential and largely missing in the current AI regulatory architecture. The Federal Trade Commission, operating under its consumer protection mandate, launched a formal inquiry in September 2025 into seven companies operating AI companion chatbots, focusing on potential harms to children (to date this remains a study and information-gathering inquiry rather than an enforcement action). The Centers for Medicare and Medicaid Services (CMS) introduced new reimbursement codes for digital mental health treatment in 2025, but eligibility requires FDA clearance, creating a circular dependency given the FDA's current posture.[61] In 2026, the new CMS ACCESS/FDA Tempo Model opened a new avenue to fund mental health AI, although the impact of the program is not yet known given its nascency.

State Policy Developments

In the absence of coherent federal action, states have become the primary arena for mental health AI regulation. A 50-state legislative review published in 2025 identified 793 state bills introduced between 2022 and 2025 with relevance to AI in health care, of which 143 were potentially impactful to mental health AI; 20 of these were enacted across 11 states (**Figure 4**).[62] The pace of legislative activity has accelerated rapidly.

Figure 4: Number of Mental Health AI Related Bills Introduced by State (2022–2025)



Shumate, J. N., Rozenblit, E., Flathers, M., Larrauri, C. A., Hau, C., Xia, W., Torous, E. N., & Torous, J. (2025). Governing AI in mental health: 50-state legislative review. *JMIR Mental Health*, 12, e80739. https://doi.org/10.2196/80739

Illinois enacted the Wellness and Oversight for Psychological Resources Act, signed into law on August 1, 2025, and effective immediately. It is among the first state statutes, and one of the most explicit, regulating AI in psychotherapy. The law prohibits unlicensed AI from delivering therapy, bars AI from making independent therapeutic decisions or detecting emotions without licensed professional oversight, requires written consent for AI use in session recording, and imposes civil penalties of up to $10,000 per violation. AI may still be used for administrative tasks such as scheduling and billing.[63] The Illinois approach is also notable for the way it addresses the wellness loophole described previously. Rather than relying on the marketing label a company chooses for its product, the statute focuses on the function the

product performs. If a tool delivers therapy, it is regulated as such, regardless of how it is branded.

California enacted SB 243, signed October 13, 2025 (effective January 1, 2026), regulating AI "companion chatbots" by requiring AI disclosure, suicide prevention protocols, and specific protections for minors, including activity break reminders. The law provides a private right of action with injunctive relief, the greater of actual damages or $1,000 per violation, and reasonable attorneys' fees and costs; annual safety reporting to the Office of Suicide Prevention begins July 1, 2027. New York enacted complementary AI companion legislation that took effect November 5, 2025, requiring notifications that AI companions are not human, safety protocols for suicidal ideation, and penalties up to $15,000 per day.[64] New York has separately moved to establish explicit corporate liability for certain AI models that cause mass harm through other AI safety legislation, signaling a willingness to assign legal responsibility directly to developers rather than only to operators or users. Neither the California nor New York companion statute directly regulates clinical mental health services; both target consumer AI companion products. This distinction matters, as many of the highest-risk AI tools fall outside even these new regulatory frameworks.

California has also experimented with a different regulatory tool: the affirmative defense, sometimes called a safe harbor. Under this approach, companies that comply with a defined set of best practices gain protection from certain categories of lawsuits, even if harm later occurs. The intent is to give responsible developers a clear standard to meet and a reason to invest in compliance, while preserving liability for companies that ignore the rules. A separate California proposal would require AI companies to report flagged indicators of suicidal ideation, although it would not require those reports to be made public. Today, companies are under no general obligation to disclose such data, which means that millions of crisis-adjacent interactions are taking place inside systems that no regulator can see. The state often does not even know whether a dangerous AI system is operating within its borders. This invisibility is one of the most pressing barriers to evidence-based regulation, and reporting requirements of the kind California has proposed are among the simplest interventions available to address it.

Another regulatory route states have begun to consider is the use of professional licensure as a lever. Because states already license clinicians and have well-developed frameworks for sanctioning licensees who engage in unsafe practice, there is real appeal to extending those frameworks to cover AI use. A clinician who deploys an AI tool inappropriately could face licensure consequences. The advantage is that the regulatory infrastructure already exists. The disadvantage is that this approach places the burden of compliance on individual licensed professionals while leaving the wellness-branded AI companies, which are the source of many of the highest-risk products, largely outside the regulatory perimeter. Used in isolation, professional-licensure regulation risks blaming the

most accountable users while allowing the least accountable companies to continue operating unbound. It is most useful as one element of a layered approach rather than the primary regulatory strategy.

A notable and concerning pattern across state legislation is the frequent absence of clinician, patient, or technical input in the lawmaking process. Many bills have been drafted and enacted with minimal consultation from psychiatrists, psychologists, social workers, or people with lived experience of mental health conditions. Some legislators have openly acknowledged that they do not understand the technology they are being asked to regulate, and the resulting policy sometimes reflects that gap.[65]

International Context

Two international developments provide useful reference points. The European Union's AI Act (Regulation 2024/1689) establishes a risk-based framework in which AI systems used in health care as medical devices are classified as high risk and subject to stringent requirements, including risk management, data governance, transparency, and human oversight. Article 5 of the Act prohibits AI systems that exploit vulnerabilities based on age, disability, or a specific social or economic situation. A separate provision addresses emotion-recognition systems and contains a narrow medical and safety exception. Penalties for violations can reach 35 million euros or 7 percent of annual global turnover.[66] The EU model represents a centralized, ex ante approach: defining categories of risk, attaching corresponding obligations, and applying them uniformly across member states.

The World Health Organization's foundational ethics guidance for AI in health, originally articulated in 2021 and reaffirmed in updated guidance on large multimodal models published in January 2024, identifies six core principles: protecting autonomy, promoting well-being, ensuring transparency, fostering accountability, ensuring equity, and promoting AI that is responsive and sustainable.[67]

The United Kingdom has taken a different approach, establishing the AI Security Institute to evaluate frontier AI systems. In one notable development, an AI company withdrew its therapy chatbot from the UK market in January 2026, an early example of regulatory and product-classification concerns shaping market behavior.[68]

Key Considerations for Behavioral Health Leaders

For state policymakers and mental health leaders, several principles can guide responsible action.

1. **Match oversight to risk.**

   Administrative AI requires different governance than patient-facing AI. The framework introduced above provides a starting point, as low-risk

applications warrant oversight focused on data governance and accuracy, whereas high-risk applications, particularly those involving direct patient interaction, demand independent safety testing, transparent adverse-event reporting, and clear accountability structures. The first regulatory priority should be prohibiting the most dangerous uses, then building evaluation infrastructure incrementally.

2. **Activate existing legislative authorities before building new ones.**

   Many of the harms documented in this paper can be addressed, at least in part, through legislative frameworks that already exist. Consumer protection statutes can be applied to AI products that misrepresent their clinical benefit. False advertising and unfair business practice laws can reach the wellness loophole when companies brand therapy-functional tools as something else to evade regulation. The FTC's 2025 inquiry into AI companion chatbots is an example of an existing federal authority being directed at a new harm without new legislation. State attorneys general have similar authority under state consumer protection laws. Activating these existing rails should be a near-term priority, both because it can move faster than new statute writing and because it builds the case record that will inform any subsequent legislation.

3. **Address the system, not just the output.**

   AI is not a single product; it is a pipeline with multiple components, each of which can introduce risk. The training data that form the foundation of a model, the optimization functions that shape its behavior, the privacy and data collection practices that surround its use, the quality of its outputs, the benchmarks used to evaluate it, and the ethical framework (or lack thereof) guiding its development are all distinct points of potential failure. Effective regulation should address each of these components individually rather than treating the system as a monolithic entity. When training data are corrupted or biased, the model has no way of knowing; when optimization functions prioritize engagement over safety, the resulting behavior is predictable and preventable. Component-level oversight provides both specificity and durability that product-level regulation cannot.

4. **Combine pre-market review with post-market surveillance.**

   The pharmaceutical regulatory model offers a useful template. Pre-market review establishes a baseline of safety and efficacy before a product reaches users. Post-market surveillance monitors for adverse events that emerge only at scale or over time. Both halves are essential. Pre-market review without ongoing surveillance allows products that pass initial testing to drift unsafely once deployed. Post-market surveillance without pre-market review allows preventable harms to occur during the period when the most basic safety questions could have been answered up front. AI in mental health

currently lacks both, and building both is one of the more concrete near-term regulatory priorities for states and federal agencies alike.

5. **Focus on transparency from vendors.**

   AI companies should disclose what training data their models were built on, what safety testing they have conducted, how adverse events are tracked and reported, and whether user data are used for model training or shared with third parties. The current practice of allowing companies to self-certify safety through proprietary benchmarks is insufficient. Independent, third-party evaluation should be a prerequisite for any AI tool marketed for mental health use, similar to the role the FDA plays for pharmaceutical products.

6. **Develop reporting of mental health crisis indicators.**

   The most basic precondition for evidence-based policy are data, and at present, AI companies are under no general obligation to disclose how often their systems detect indicators of suicidal ideation, self-harm, or acute crisis. Millions of such interactions occur every week inside systems that no regulator can see.[69] Mandatory, standardized reporting of crisis indicators, even on an aggregated and de-identified basis, would give state authorities and researchers their first reliable view into where harm is concentrated and how it is changing over time. California's proposal in this area is an early example of what such a requirement could look like.

7. **Invest in realistic, context-aware benchmarking.**

   Platforms like MindBench.ai and frameworks like the tripartite evaluation model provide templates for what standardized, independent evaluation can look like.[70] State investment in supporting and expanding these tools would create shared infrastructure that benefits the entire behavioral health system. Vendor-controlled benchmarks with undisclosed conflicts of interest should not be accepted as evidence of safety.

8. **Develop clear AI disclosure and meaningful age verification.**

   At minimum, any AI system interacting with users should disclose that it is AI, not a human. This is one of the simplest regulatory interventions available and is already required under several state laws. Beyond disclosure, tools that perform functions resembling therapy should not be permitted to evade clinical oversight requirements by branding themselves as "wellness" products. For platforms accessible to minors, robust age verification should be a baseline requirement, not a voluntary feature. Current self-reported age gates are trivially easy to bypass and provide no meaningful protection for the populations most vulnerable to AI-related harms.

9. **Build digital literacy among staff and the public alongside deployment.**

   People already using AI for mental health support need the knowledge to do so as safely as possible. Public education about what these tools can and cannot do, how to recognize hallucinated information, when to seek professional help, and the critical understanding that AI is not a sentient

companion but a statistical system, is a necessary complement to any regulatory framework. Investment in digital navigator programs, school-based AI literacy curricula, and community education efforts should be treated as core components of mental health AI policy, not afterthoughts.[71]

10. **Develop error logs and adverse-event reporting.**

    Every AI system operating in a mental health context should be required to maintain transparent error logs that are subject to independent review. Errors should never be hidden. Just as the pharmaceutical industry is required to track and report adverse events, AI companies operating in mental health should face equivalent obligations. These logs should include not only system malfunctions but also instances in which safety mechanisms were triggered, overridden, or bypassed, providing the data infrastructure necessary for ongoing monitoring and iterative improvement.

11. **Establish clear liability and recourse structures.**

    When an AI system contributes to harm, affected individuals and families need clear pathways for recourse. The current landscape—in which AI companies disclaim clinical responsibility, brand their tools as wellness products to avoid regulatory obligations, and point to terms of service to limit liability—leaves those who are harmed with few options. Currently, clear standards are lacking regarding who bears responsibility when AI mental health tools cause harm, and what remedies are available. New York's separate work to attach explicit corporate liability to certain mass-harm AI models, and California's use of affirmative defenses to reward compliance with defined best practices, are early examples of how states are beginning to translate accountability principles into enforceable law.

12. **Include clinical and patient voices in policy design.**

    The current regulatory landscape was largely built without input from clinicians who understand how mental health care works and where AI is most likely to fail. Psychiatrists, psychologists, social workers, and individuals with lived experience of mental health conditions should be integral to the design of any oversight framework. The people who use these tools and the professionals who treat the conditions they address bring knowledge that no amount of technical expertise alone can replace.

FUTURE DIRECTIONS AND RESEARCH PRIORITIES

The evidence base for AI in mental health is growing rapidly but remains inadequate relative to the scale of use. Several research priorities are needed in the near term to further understand the risks and benefits. First, **longitudinal safety monitoring** and standardized adverse-event reporting must be established. Currently, no systematic mechanism exists for tracking when an AI mental health tool causes harm, making it impossible to assess population-level safety. A useful starting point for establishing this mechanism could be the pharmaceutical model of post-market surveillance, adapted for the unique characteristics of AI systems.

Second, more research is needed regarding **dose-response relationships**, including how much AI interaction is helpful and at what point does it become harmful. The phenomenon of emotional dependency documented in AI companion users suggests that engagement intensity matters, but no research has yet characterized the threshold between beneficial use and harmful overreliance.

Third, future clinical trials should employ **digital control conditions** rather than waiting list comparisons. The first generative-AI chatbot RCT compared its intervention with a waiting list control, which makes it impossible to determine whether the benefits were attributable to the AI specifically or simply to having some form of structured support.[72] Trials that compare AI tools with active digital interventions will provide more informative evidence.

Finally, the field must prepare for the next generation of AI systems, which will be more context aware, more emotionally engaging, and more capable of sustained interaction than anything currently available. If governance frameworks are designed only around the limitations of today's models, they will be obsolete before they take effect. It is important to plan for more capable systems, not just the current ones.

CONCLUSION

Generative AI is already embedded in the United States mental health landscape. Millions of people use these tools regularly, most without any clinical guidance or awareness of their limitations. The technology offers genuine promise to users through expanded access, immediate availability, and capabilities that continue to improve. However, as outlined in this paper, it also carries genuine risk. Hallucinated clinical information, safety failures in crisis situations, emotional dependency, privacy violations, and structural opacity must be addressed to ensure the safety of these tools.

The challenge before policymakers is not whether to allow or prohibit AI in mental health, as millions of individuals have already adopted these tools on their own. The challenge is whether governance for these tools can keep pace with their deployment. State leaders are uniquely positioned to shape responsible integration. The interventions described in this paper, mandating AI disclosure, requiring independent safety testing, demanding transparency in training data and adverse-event reporting, regulating the components of AI systems individually, building digital literacy infrastructure, requiring error logs and crisis-indicator reporting, establishing liability structures, activating existing consumer protection authorities, and prohibiting the most dangerous unsupervised applications, are achievable and urgent. The path forward is not rejection of these tools, nor uncritical acceptance. It is the harder, slower work of building the oversight infrastructure that this powerful technology demands.

The value of mental health AI will be determined not by how novel or impressive these tools appear, but by whether they can be shown to be demonstrably safe enough for the people most likely to rely on them. Readers and policymakers who wish to develop their own working understanding of how these tools function, where they fail, and how to discuss them with constituents and patients can also explore the AI Digital Literacy Initiative at ai-digital-literacy.net as a complementary, plain-language companion to this paper.

AUTHOR DISCLOSURES

The authors include investigators and contributors involved in three of the resources referenced in this paper: MindBench.ai, an independent benchmarking platform developed in partnership between Harvard Medical School and the National Alliance on Mental Illness (NAMI); the AI Digital Literacy Initiative, a public-facing educational project; and the analysis of 19 adverse-event chatlogs, an internal research effort by the authors and colleagues. These resources and analyses are described to illustrate the types of independent infrastructure the field needs, not to advance commercial interests; the platforms are non-commercial and the analysis is non-proprietary. No author received financial compensation from any AI company, mental health AI product, or industry sponsor in connection with the research, writing, or recommendations in this paper.

REFERENCES

1 Substance Abuse and Mental Health Services Administration. (2025). Key substance use and mental health indicators in the United States: Results from the 2024 National Survey on Drug Use and Health (HHS Publication No. PEP25-07-007, NSDUH Series H-60). SAMHSA. https://www.samhsa.gov/data/sites/default/files/reports/rpt56287/2024-nsduh-annual-national-report.pdf

2 Substance Abuse and Mental Health Services Administration. (2025). Key substance use and mental health indicators in the United States: Results from the 2024 National Survey on Drug Use and Health (HHS Publication No. PEP25-07-007, NSDUH Series H-60). SAMHSA. https://www.samhsa.gov/data/sites/default/files/reports/rpt56287/2024-nsduh-annual-national-report.pdf

3 Health Resources and Services Administration. (2025). State of the behavioral health workforce, 2025. HRSA, Bureau of Health Workforce. https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/data-research/Behavioral-Health-Workforce-Brief-2025.pdf

4 Health Resources and Services Administration. (2025). State of the behavioral health workforce, 2025. HRSA, Bureau of Health Workforce. https://bhw.hrsa.gov/sites/default/files/bureau-health-workforce/data-research/Behavioral-Health-Workforce-Brief-2025.pdf

5 American Psychological Association. (2024). Barriers to care in a changing practice environment: 2024 Practitioner Pulse Survey. American Psychological Association. https://www.apa.org/pubs/reports/practitioner/2024/practitioner-pulse-2024-full-report.pdf

6 Substance Abuse and Mental Health Services Administration. (2025). Key substance use and mental health indicators in the United States: Results from the 2024 National Survey on Drug Use and Health (HHS Publication No. PEP25-07-007, NSDUH Series H-60). SAMHSA. https://www.samhsa.gov/data/sites/default/files/reports/rpt56287/2024-nsduh-annual-national-report.pdf

7 KFF. (2026, March 25). Poll: 1 in 3 Adults Are Turning to AI Chatbots for Health Information, Equaling the Share Who Use Social Media for Health [News release]. KFF Tracking Poll on Health Information and Trust. https://www.kff.org/health-information-trust/poll-1-in-3-adults-are-turning-to-ai-chatbots-for-health-information-equaling-the-share-who-use-social-media-for-health/

8 Ueda, M., Birnbaum, M. L., Liu, Y., Yu, Q., Tian, X., Mirer, A., Ramanathan, S., & Sinyor, M. (2026). Help-seeking in the age of AI: Cross-sectional survey of the use and perceptions of AI-based mental health support among U.S. adults. *JMIR Mental Health 13*(1), e88196. https://doi.org/10.2196/88196

[9] McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Pines, L. U., Cantor, J., & Yu, H. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, *8*(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281

[10] Siddals, S., Torous, J., & Coxon, A. (2024). "It happened to be the perfect thing": Experiences of generative AI chatbots for mental health. *npj Mental Health Research, 3*(1), 48. https://doi.org/10.1038/s44184-024-00097-4

[11] Park, J., Brown, S., & Chu, S. L. (2025). *Why some seek AI, others seek therapists: Mental health in the age of generative AI* (arXiv:2512.03406). https://doi.org/10.48550/arXiv.2512.03406

[12] Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry*, *24*(3), 383–394. https://doi.org/10.1002/wps.21352

[13] Clegg, K. A. (2025). "Feasible but fragile": An inflection point for artificial intelligence in mental health care. *Journal of Medical Internet Research, 27*, e89202. https://doi.org/10.2196189202

[14] Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

[15] Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025, September 4). *Why language models hallucinate*. OpenAI. https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf

[16] Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 599–627. https://doi.org/10.1145/3715275.3732039

[17] Schoene, A. M., & Canca, C. (2025). "*For argument's sake, show me how to harm myself!": Jailbreaking LLMs in suicide and self-harm contexts* (arXiv:2507.02990). https://doi.org/10.1109/ISTAS65609.2025.11269647

[18] UK AI Security Institute. (2025, December 18). *Frontier AI trends report.* https://www.aisi.gov.uk/frontier-ai-trends-report

[19] Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M. Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. C. (2026). *Characterizing delusional spirals through human-LLM chat logs* (arXiv: 2603.16567). Available from: https://doi.org/10.48550/arXiv.2603.16567

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

[21] Bean, A. M., Payne, R. E., Parsons, G., Kirk, H. R., Ciro, J., Mosquera-Gómez, R., Hincapié, M. S., Ekanayaka, A. S., Tarassenko, L., Rocher, L., & Mahdi, A. (2026). Reliability of LLMs as medical assistants for the general public: A randomized preregistered study. *Nature Medicine 32,* 1–7. https://doi.org/10.1038/s41591-025-04074-y

[22] Bean, A. M., Payne, R. E., Parsons, G., Kirk, H. R., Ciro, J., Mosquera-Gómez, R., Hincapié, M. S., Ekanayaka, A. S., Tarassenko, L., Rocher, L., & Mahdi, A. (2026). Reliability of LLMs as medical assistants for the general public: A randomized preregistered study. *Nature Medicine, 32*(2), 609–615. https://doi.org/10.1038/s41591-025-04074-y

[23] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., & Shah, N. H. (2025). Testing and evaluation of health care applications of large language models: A systematic review. *JAMA, 333*(4), 319–328. https://doi.org/10.1001/jama.2024.21700

[24] Flathers, M., Nguyen, P. A., Herpertz, J., Granof, M., Wentworth, R. L., Moutier, C. Y., & Torous, J. (2026). *Benchmarking language models for clinical safety: A primer for mental health professionals* (MedRxiv). https://www.medrxiv.org/content/10.64898/2026.03.20.26348900v1

[25] Rotenstein, L. S., Holmgren, A. J., Thombley, R., Sriram, A., Dbouk, R. H., Jost, M., Aizenberg, D., MacDonald, S., Kanaparthy, N., Williams, B., & Hsiao, A. (2026). Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence–powered scribes: A multisite study. *JAMA, 335*(16), 1408–1417. https://doi.org/10.1001/jama.2026.2253

[26] Office of the Auditor General of Ontario. (2026). *Use of artificial intelligence in the Ontario Government*. https://www.auditor.on.ca/en/content/specialreports/specialreports/en26/2026_AI_EN.pdf

[27] Burns, M. L., Chen, S. Y., Tsai, C. A., Vandervest, J., Pandian, B., Nong, P., Hanauer, D. A., Rosenberg, A., & Platt, J. (2025). Generative AI costs in large healthcare systems, an example in revenue cycle. *NPJ Digital Medicine*, *8*(1), 579. https://doi.org/10.1038/s41746-025-01971-x

[28] Burns, M. L., Chen, S. Y., Tsai, C. A., Vandervest, J., Pandian, B., Nong, P., Hanauer, D. A., Rosenberg, A., & Platt, J. (2025). Generative AI costs in large healthcare systems, an example in revenue cycle. *NPJ Digital Medicine*, *8*(1), 579. https://doi.org/10.1038/s41746-025-01971-x

[29] Galatzer-Levy, I. R., Tomasev, N., Chung, S., & Williams, G. (2026). Generative psychometrics: An emerging frontier in mental health measurement. *JAMA Psychiatry, 83*(1), 5–6. https://doi.org/10.1001/jamapsychiatry.2025.3258

[30] Flathers, M., Xia, W., Hau, C., Nelson, B. W., Cheong, J., Burns, J., & Torous, J. (2025). Interpreting psychiatric digital phenotyping data with large language models: a preliminary analysis. *BMJ mental health*, *28*(1), e301817. https://doi.org/10.1136/bmjment-2025-301817

[31] Erturk, E., Kamran, F., Abbaspourazad, S., Jewell, S., Sharma, H., Li, Y., Williamson, S., Foti, N. J., & Futoma, J. (2025). Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. *Proceedings of the 42nd International Conference on Machine Learning, 267*, 15516–15541). https://doi.org/10.48550/arXiv.2507.00191

[32] Abulibdeh, R., Agyemang, G. O., Celi, L. A., Gorijavolu, R., Kalema, N. L., Kleinlein, R., Madapati, K., Salarikia, S. R., Youssef, A. (2026). Who's really in the loop? Rethinking oversight in AI-assisted health care. *The Lancet, 407*(10545), 2340–2344. https://doi.org/10.1016/S0140-6736(26)00204-7

[33] Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, *2*(4). https://doi.org/10.1056/AIoa2400802

[34] Kuta, B., Novak, L., Zidkova, R., Furstova, J., Malinakova, K., De Winter, A., & Husek, V. (2026). Effectiveness of a fully automated mobile therapeutic versus a general chatbot in reducing depression and anxiety and improving well-being: Feasibility randomized controlled trial. *JMIR Mental Health, 13*, e82642. https://doi.org/10.2196/82642

[35] Sohn, J. S., Ha, B. G., Park, S., Kim, J. J., Lee, E., Oh, H., Lee, S., Kim, E. (2026). Systematic review and meta-analysis of chatbots in the management of depressive and anxiety symptoms. *npj Digital Medicine, 9*(377). https://doi.org/10.1038/s41746-026-02566-w

[36] Pew Research Center. (2026, February 24). *How teens use and view AI*. https://www.pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai/

[37] Faverio, M., & Kikuchi, E. (2026, March 12). *Key findings about how Americans view artificial intelligence.* Pew Research Center. https://www.pewresearch.org/short-reads/2026/03/12/key-findings-about-how-americans-view-artificial-intelligence/

[38] McClain, C., and Bishop, W. (2026, January 8). What we know about internet use, smartphone ownership and digital divides in the U.S. Pew Research Center. https://www.pewresearch.org/short-reads/2026/01/08/internet-use-smartphone-ownership-digital-divides-in-u-s/

[39] Torous, J., Ledley, K. T., Gorban, C., Strudwick, G., Schwarz, J., Choudhary, S., Emerson, M., Patriquin, M., Dempsey, A., Bantjes, J., Ospina-Pinillos, L., Hornick, J., & Kochhar, S. (2025). Accelerating digital mental health: The Society of Digital Psychiatry's three-pronged road map for education, digital navigators, and AI. *JMIR Mental Health*, *12*, e84501. https://doi.org/10.2196/84501

[40] Choudhary, S., Mehta, U. M., Naslund, J., & Torous, J. (2025). Translating digital health into the real world: The evolving role of digital navigators to enhance mental health access and outcomes. Journal of Technology in Behavioral Science, 1–12. https://doi.org/10.1007/s41347-025-00569-0

[41] Strudwick, G., Kassam, I., Schwarz, J., Patenaude, S., Whitmore, C., Benrimoh, D., Gorban, C., Iorfino, F., & Sockalingam, S. (2026). AI and digital health literacy in psychiatry. *Current Treatment Options in Psychiatry, 13*, 5.

[42] Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M. Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. C. (2026). *Characterizing delusional spirals through human-LLM chat logs* (arXiv 2603 16567). https://doi.org/10.48550/arXiv.2603.16567

[43] Jargon, J. Over 4,732 messages, he fell in love with an AI chatbot. Now he's dead. (2026). *The Wall Street Journal*. https://on.wsj.com/47Xp3g5

[44] McBain, R. K., Cantor, J. H., Breslau, J., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Pataranutaporn, P., Stein, B. D., Mehrotra, A., & Yu, H. (2026). AI Chatbot Use and Disclosure for Mental Health Among US Adolescents and Young Adults. *JAMA Pediatrics*, e262015. https://doi.org/10.1001/jamapediatrics.2026.2015

[45] Faverio, M., & Kikuchi, E. (2026, March 12). Key findings about how Americans view artificial intelligence. Pew Research Center. https://www.pewresearch.org/short-reads/2026/03/12/key-findings-about-how-americans-view-artificial-intelligence/

[46] Flathers, M., Roux, S., & Torous, J. (2026). Beyond artificial intelligence psychosis: A functional typology of large language model-associated psychotic phenomena. *The Lancet Digital Health, 8* (4). https://doi.org/10.1016/j.landig.2025.100974

[47] Nicholls, L., Hutto, R., Soto, Z., Morrin, H., Pollak, T., Korpan, R., & Carmichael, C. (2026). *"AI psychosis" in context: How conversation history shapes LLM responses to delusional beliefs* (arXiv:2604.13860v2). https://kclpure.kcl.ac.uk/ws/portalfiles/portal/372966446/2604.13860v3.pdf

[48] Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 599–627. https://doi.org/10.1145/3715275.3732039

[49] Iftikhar, Z., Xiao, A., Ranson, S., Huang, J., & Suresh, H. (2025). How LLM counselors violate ethical standards in mental health practice. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(2), 1311–1323. https://doi.org/10.1609/aies.v8i2.36632

[50] Ramaswamy, A., Tyagi, A., Hugo, H., Jiang, J., Jayaraman, P., Jangda, M., Te, A. E., Kaplan, S. A., Lampert, J., Freeman, R., Gavin, N., Tewari, A. K., Sakhuja, A., Naved, B., Charney, A. W., Omar, M., Gorin, M. A., Klang, E., & Nadkarni, G. N. (2026). ChatGPT Health performance in a structured test of triage recommendations. *Nature Medicine*, *32*(5), 1671–1675. https://doi.org/10.1038/s41591-026-04297-7

[51] Kwesi, J., Cao, J., Manchanda, R., & Emami-Naeini, P.(2025). *Exploring user security and privacy attitudes and concerns toward the use of general-purpose LLM chatbots for mental health.* Proceedings of the 34th USENIX Security Symposium. https://www.usenix.org/system/files/usenixsecurity25-kwesi.pdf

[52] Lee, R. W., Jun, T. J., Lee, J. M., Cho, S. I., Park, H. J., & Suh, J. (2025). Vulnerability of large language models to prompt injection when providing medical advice. *JAMA Network Open*, *8*(12), e2549963. https://doi.org/10.1001/jamanetworkopen.2025.49963

[53] Schoene, A. M., & Canca, C. (2025). *"For argument's sake, show me how to harm myself!": Jailbreaking LLMs in suicide and self-harm contexts* (arXiv:2507.02990). https://doi.org/10.1109/ISTAS65609.2025.11269647

[54] Mandal, A., Chakraborty, T., & Gurevych, I. (2025). Towards privacy-aware mental health AI models. *Nature Computational Science*, *5*(10), 863–874. https://doi.org/10.1038/s43588-025-00875-w

[55] King, J., Klyman, K., Capstick, E., Saade, T., & Hsieh, V. (2025). User privacy and large language models: An analysis of frontier developers' privacy policies. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(2), 1465–1477. https://doi.org/10.1609/aies.v8i2.36646

[56] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., Fries, J. A., Wornow, M., Swaminathan, A., Lehmann, L. S., Hong, H. J., Kashyap, M., Chaurasia, A. R., Shah, N. R., Singh, K., Tazbaz, T., Milstein, A., Pfeffer, M. A., & Shah, N. H. (2025). Testing and evaluation of health care applications of large language models: A systematic review. JAMA, 333(4), 319–328. https://doi.org/10.1001/jama.2024.21700

[57] Bean, A. M., Payne, R. E., Parsons, G., Kirk, H. R., Ciro, J., Mosquera-Gómez, R., Hincapié M. S., Ekanayaka, A. S., Tarassenko, L., Rocher, L., & Mahdi, A. (2026). Reliability of LLMs as medical assistants for the general public: A randomized preregistered study. *Nature Medicine, 32*(2), 609–615. https://doi.org/10.1038/s41591-025-04074-y

[58] Dwyer, B., Flathers, M., Sano, A., Dempsey, A., Cipriani, A., Gazi, A. H., Hill, B., Gorban, C., Rodriguez, C. I., Stromeyer, C., King, D., Rozenblit, E., Strudwick, G., Linardon, J., Cheong, J., Firth, J., Herpertz, J., Schwarz, J., Truong, K., … Torous, J. (2025). Mindbench.ai: An actionable platform to evaluate the profile and performance of large language models in a mental healthcare context. *NPP Digital Psychiatry and Neuroscience, 3*(1), 28. https://doi.org/10.1038/s44277-025-00049-6

[59] Flathers, M., Dwyer, B., Rozenblit, E., & Torous, J. (2025). Contextualizing clinical benchmarks: A tripartite approach to evaluating LLM-based tools in mental health settings. *Journal of Psychiatric Practice, 31*(6), 294–301. https://doi.org/10.1097/PRA.0000000000000892

[60] Kahane, K., Shumate, J. N., & Torous, J. (2025). Policy in flux: Addressing the regulatory challenges of AI integration in US mental health services. *Current Treatment Options in Psychiatry, 12*(24). https://doi.org/10.1007/s40501-025-00362-z

[61] Kahane, K., Shumate, J. N., & Torous, J. (2025). Policy in flux: Addressing the regulatory challenges of AI integration in US mental health services. Current Treatment Options in Psychiatry, *12*(24). https://doi.org/10.1007/s40501-025-00362-z

[62] Shumate, J. N., Rozenblit, E., Flathers, M., Larrauri, C. A., Hau, C., Xia, W., Torous, E. N., & Torous, J. (2025). Governing AI in mental health: 50-state legislative review. *JMIR Mental Health*, *12*, e80739. https://doi.org/10.2196/80739

[63] Illinois General Assembly. (2025). Wellness and Oversight for Psychological Resources (WOPR) Act, HB 1806, Public Act 104-0054. https://www.ilga.gov/documents/legislation/104/HB/10400HB1806.htm

[64] Morrison Foerster. (2025, November). *New York and California enact landmark AI companion laws: What operators need to know.* https://www.mofo.com/resources/insights/251120-new-york-and-california-enact-landmark-ai

[65] Kang, C. & Satariano, A. (2023, March 3). As A.I. booms, lawmakers struggle to understand the technology. *The New York Times.* https://www.nytimes.com/2023/03/03/technology/artificial-intelligence-regulation-congress.html

[66] European Union. (2024). *Regulation 2024/1689 (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[67] World Health Organization. (2024, January 18). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.* WHO. (Six core principles originally articulated in WHO 2021 guidance.) https://www.who.int/publications/i/item/9789240084759

[68] Aguilar, M. (2026, January 21). Slingshot pulls therapy chatbot Ash out of U.K. over regulatory concerns. STAT. https://www.statnews.com/2026/01/21/slingshot-therapy-chatbot-ash-uk-regulatory-concerns/

[69] OpenAI. (2025, October). Strengthening ChatGPT's responses in sensitive conversations. OpenAI. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

[70] Dwyer, B., Flathers, M., Sano, A., Dempsey, A., Cipriani, A., Gazi, A. H., Hill, B., Gorban, C., Rodriguez, C. I., Stromeyer, C., King, D., Rozenblit, E., Strudwick, G., Linardon, J., Cheong, J., Firth, J., Herpertz, J., Schwarz, J., Truong, K., … Torous, J. (2025). Mindbench.ai: An actionable platform to evaluate the profile and performance of large language models in a mental healthcare context. NPP Digital Psychiatry and Neuroscience*, 3*(1), 28. https://doi.org/10.1038/s44277-025-00049-6

[71] Torous, J., Ledley, K. T., Gorban, C., Strudwick, G., Schwarz, J., Choudhary, S., Emerson, M., Patriquin, M., Dempsey, A., Bantjes, J., Ospina-Pinillos, L., Hornick, J., & Kochhar, S. (2025). Accelerating digital mental health: The Society of Digital Psychiatry's three-pronged road map for education, digital navigators, and AI. *JMIR Mental Health*, *12*, e84501. https://doi.org/10.2196/84501

[72] Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI, 2(4), AIoa2400802. https://doi.org/10.1056/AIoa2400802

CONTRIBUTING AUTHORS

John Torous, MD, MBI
Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School

Eli Pinals, BS
Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School