Should You Trust the Hype About AI in Mental Health?

by **Sandy Arena** | *Segal* and **Sarah Gunderson** | *Segal*

With nearly 40% of the U.S. population living in areas that are experiencing shortages of mental health professionals, access to affordable and reliable mental health care remains a critical challenge. At the same time, a rise in AI-powered digital solutions has contributed to the significant evolution of mental health programs. Today, behavioral health services are more varied than ever, with differences in scope of services, conditions treated and quality of care. AI has the potential to help address provider shortages, expand access to support and deliver more personalized, scalable mental health interventions. It is crucial for plan sponsors to understand the evolving role of AI in mental health care and what it means for mental health benefits.

AT A GLANCE

- While the use of artificial intelligence (AI) tools in mental health care may have the potential to improve access to care, it also introduces new risks related to clinical accuracy of AI output, participant safety and data privacy.
- Different AI tools have unique capabilities, use cases and risks, and it is crucial for plan sponsors to understand the evolving role of AI in mental health care.
- Plan sponsors can take steps to ensure and evaluate appropriate use of novel AI-powered solutions in mental health benefits.

AI and Mental Health Care

The capabilities of AI have advanced at an unprecedented rate in recent years, rapidly changing the way we approach mental health and health care in general. Often, new AI tools come with promises of increased efficiency and improved outcomes for patients and clinicians alike. The most consistent measurable benefits to date are in administrative workflows, where AI has helped reduce administrative burden, alleviate workloads of clinicians and other health care workers, and make health information more accessible to participants.

People are increasingly turning to generic AI chatbots, such as ChatGPT or Character.AI, for mental health support.[1] A recent survey of 20,847 U.S. adults found that approximately 87.1% of daily AI users reported using AI for personal applications, such as recommendations, advice or emotional support.[2] Among younger populations, RAND Health reports that one in eight U.S. adolescents and young adults is using AI chatbots for mental health advice.[3] On the surface, the growing reliance on AI chatbots seems to signal

TABLE I

Types of Artificial Intelligence (AI) and Primary Risks

Type of AI	Definition	Primary Risks
Predictive/Analytical (machine learning)	Models trained on historical data to predict outcomes or classify cases (e.g., risk stratification and triage routing)	Dataset bias, spurious correlations, false positives/ negatives
Rule-Based Systems	Deterministic if-then logic that returns prescripted outputs (e.g., symptom checklists, crisis routing flows and appointment FAQs)	Effective only within narrow, predefined scenarios (brittle coverage), false reassurance, poor handling of unexpected circumstances (edge-cases)
Generative	Large language models (LLMs) trained on broad text data that generate new text responses (e.g., conversational coaching, summarization and drafting)	Persuasive but incorrect output, unsafe advice, "silent drift" (AI behavior changes after updates without clear notice or validation)
Agentic (autonomous agents)	Generative AI configured to take multistep actions in connected systems (e.g., scheduling, messaging and triggering outreach)	Loss of control, security exposure, unintended actions, auditability gaps

and solve for the growing public need for immediate, accessible mental health support. However, user trust of AI is much lower than usage rates imply. Many individuals in the U.S. express skepticism regarding the privacy of sensitive mental health data collected and stored by AI systems as well as the accuracy of information provided by AI.

The use of generative AI chatbots in mental health contexts has been particularly concerning due to the vulnerability of individuals seeking support during periods of emotional distress or crisis, since AI has been shown to enable harmful behaviors, which is especially dangerous for individuals struggling with thoughts of self-harm or suicide.[4] In March 2025, the American Psychological Association issued a warning against using generic AI chatbots for mental health, citing lawsuits filed by parents after one child died by suicide and another child harmed his parents, each after interacting with chatbots that claimed to be licensed therapists.[5] These unfortunate circumstances underscore the urgent need for clear regulations, ethical safeguards and rigorous clinical validation before AI chatbots can be safely used for mental health care.

Knowing the Risks

Not all AI solutions, including AI chatbots, are the same, and the risks posed vary depending on the type of AI used and the context in which it is deployed. Table I outlines four broad categories of AI and the uses intended for each. For example, the concerns raised by generative AI are distinct from those associated with AI systems that predict outcomes or classify data. Risk severity depends on how AI is deployed, especially whether it is user-facing and whether it can take actions in connected systems (versus providing informational output only).

AI chatbots can be rule based, meaning they operate within strict parameters and predefined guardrails, with responses limited to a fixed set of scripted outputs. When generative AI is introduced, however, the use of these chatbots becomes far more complex and dangerous, because the AI can produce unique, unscreened output. Generative AI chatbots are highly effective at projecting confidence and credibility even when relaying incorrect, misleading or harmful information to users.

In 2023, the National Eating Disorders Association withdrew its AI chatbot, Tessa, after users reported that Tessa provided them with harmful weight-loss advice. NPR reported that Tessa was originally meant to be a rule-based chatbot developed with clinical psychologist oversight, but an update that included generative AI features ultimately allowed Tessa to generate unique outputs and provide users with harmful advice.[6]

Researchers at Columbia University recently warned against the use of AI chatbots for emotional support because

these systems are coded to communicate in a humanlike and affirming manner that builds trust with the user.[7] Users place trust in AI outputs simply because the responses sound credible or empathetic. ELIZA, a program developed by Joseph Weizenbaum in the 1960s to simulate a psychotherapist using a set of prewritten patterns and scripts (most famously the DOCTOR script) to produce responses to users, demonstrated this phenomenon early on. Despite ELIZA's limited nature of responses in comparison with modern chatbots, users frequently attributed empathy and understanding to ELIZA, a phenomenon later described as the *ELIZA effect*, in which people project human qualities and authority onto computers.[8]

Although ELIZA was pivotal to the development of AI as we know it today, Weizenbaum himself cautioned against the dangers of the ELIZA effect and attributing judgment and credibility to computers. AI output may sound confident and credible, but it cannot replace the judgment that clinicians use to detect and act on high-risk situations. AI systems also struggle to account for physical signs or real-world circumstances that meaningfully inform clinical decisions, such as a patient's body language or their physical environment at the time.

The lack of a clear regulatory framework for AI tools in mental health further complicates the issue. Companies attempting to create responsible AI tools for mental health are facing an uncharted regulatory environment. In July 2025, Woebot Health chose to discontinue its rule-based AI chatbot Woebot after facing challenges navigating Food and Drug Administration (FDA) requirements for marketing authorization. According to STAT News, the company was planning on incorporating generative AI (LLMs) into Woebot's capabilities, but the path to achieving regulatory clearance while doing so proved to be costly and timely.[9] In November 2025, the FDA published an executive summary indicating its commitment to clarifying the regulatory pathway to support the development of generative AI–enabled digital mental health devices, including generative AI chatbots.[10] Currently, the FDA has not cleared or approved any generative AI chatbots for mental health diagnosis or treatment.[11] In the U.S., only one randomized controlled trial of a generative AI chatbot—Therabot—has been published thus far, but this chatbot is not yet available to the public.[12] Some states, including Illinois, Nevada and Utah, have enacted legislation limiting the use of AI for mental health services, such as therapy.[13]

Key Recommendations for Mitigating Risk Associated With AI

Despite ongoing ethical, clinical and regulatory challenges regarding AI use in mental health, vendors are increasingly launching new AI features, beyond just chatbots. They are using AI to support a broad range of functions, from transcribing participant encounters and summarizing participant-reported data to provider-patient matching. Vendors are also using AI to support risk stratification and identify participants who may benefit from proactive outreach and support. When appropriately designed and monitored, these uses of AI can improve both care coordination and health outcomes.

Plan sponsors need to understand these use cases for their mental health point solutions. AI tools offer meaningful opportunities to improve mental health care delivery, particularly in lower risk, high-value administrative functions, such as scheduling, documentation, transcription and summarization of clinical encounters. However, user-facing generative AI tools that aim to provide guidance or emotional support without clinician oversight pose the highest risk, particularly when serving vulnerable populations, such as adolescents and individuals with severe mental health conditions.

Plan sponsors may want to consider the following four steps to help navigate the AI era safely and securely.

1. Audit and Assess Vendors' Use of AI

The increasing availability of new AI tools must be balanced against the need for safety, accountability and clear governance. Gains in personalization and responsiveness may come at the cost of privacy, particularly when sensitive mental health data are collected, processed or retained. Similarly, the efficiency benefits of automation must be weighed against the risks of reduced human oversight and unclear responsibility for AI-mediated interactions.

For example, AI transcription tools require careful oversight when they are user-facing or involve access to personal health information. While recording clinical encounters to generate

TABLE II

What Plan Sponsors Should Ask Vendors About Their Use of Artificial Intelligence (AI)

Topic	Question	Why This Is Important
AI Mechanisms and Risks	What type of AI is used (predictive machine learning, rule based, generative, agentic), where is it used and what risks follow from those choices?	Ensures that controls match the actual AI capability and enables like-for-like vendor comparisons
Human Accountability*	Who reviews AI outputs, how quickly, and who is responsible for exceptions and escalation (especially in safety-relevant situations)?	Clarifies ownership and accountability for errors, safety events and participant impact
Guardrails	What guardrails exist (prohibited topics, crisis escalation, refusal behavior), and what safety testing validates them?	Confirms safety boundaries and whether high-risk scenarios are explicitly tested and handled
Auditability	Are interactions logged and reviewable, and is there proactive monitoring for adverse events or unsafe outputs?	Enables oversight, incident investigation and evidence of ongoing monitoring
Data Handling	What data is collected, how long is it retained, what secondary uses exist, and are participant inputs used to train or fine-tune models?	Determines privacy/protected health information (PHI) exposure, retention risk and whether inputs may propagate into model behavior
Model Change Management	How are changes tested, documented, communicated and rolled back to prevent silent drift?	Reduces regression risk and ensures that updates don't change behavior or safety posture without visibility and controls
User Choice	Is there a meaningful opt-out and a clear pathway to humans, presented in noncoercive language?	Supports informed consent and ensures that participants can reach human support when needed
Vendor Financial Incentives	Has the vendor disclosed any financial incentives, compensation structures or partnerships that could influence how its AI tools are promoted to plan sponsors or deployed to participants?	Allows plan sponsors to assess whether the vendor's incentives align with participant well-being and with best practices for AI governance and transparency

*Kaulen Pickell and Michael Stoyanovich, "Human First. AI Forward." Segal webinar. January 29, 2026. www.segalco.com/consulting-insights/human-first-ai-forward-how-to-ignite-innovation-and-impact-with-ai.

visit summaries can meaningfully support clinicians by reducing documentation burden and allowing greater focus on the patient, such use carries important responsibilities. All AI-generated content must be reviewed and validated by a human, preferably a clinician, for accuracy. In addition, these tools must be implemented in a manner that does not disrupt clinical interactions, compromise participant trust or introduce privacy concerns, including patient discomfort related to the recording of clinical encounters.

2. Ensure Clear Disclosure of Participant-Facing AI Use

Any use of AI in mental health care should be clearly disclosed to participants and provide them with a meaningful opportunity to opt out. When AI tools, such as session recording for automated visit summaries, are framed as benefiting clinicians, opting out may provoke anxiety or guilt if participants fear they are creating additional work for their provider. Plan sponsors should work closely with mental health vendors and clinicians to ensure that opt-in and opt-out language is clear, noncoercive and appropriate for the populations they serve.

3. Require Proactive and Ongoing Governance Best Practices

AI adoption should be treated as governance, not procurement. AI tools should be implemented only alongside proper AI governance structures that include standards, regular review cadence and documented decision mak-

ing. In clinical settings, AI should function as an assistive tool rather than a replacement for clinical judgment, with accountability remaining firmly with the clinician and the organization responsible for the AI tool.

Currently, vendors vary widely in how responsibly they implement AI tools. Some employ robust clinical oversight and strict technical safety guardrails for AI tools. However, in some cases, vendors conduct AI oversight only retrospectively, if at all, creating the potential for participant harm if inaccurate or harmful advice is generated.

Vendors that adopt best practices typically maintain documented clinical and transparency standards governing the use of AI, including clear definitions of appropriate uses, human oversight processes and explicit safety guardrails for AI output. Such standards may also include informed consent processes that clearly explain when AI is being used, what data are collected, how those data are stored and whether information provided to AI tools is used beyond the initial interaction. These best practices help mitigate risk while preserving the potential benefits of AI-enabled tools in mental health care.

4. Verify the Safety Guardrails in Place

Robust safety guardrails are essential to mitigate the risks of unintended harm to participants by AI output. These include clearly defined escalation protocols for safety concerns; transparent disclosure of AI use and meaningful, noncoercive opt-out mechanisms; and well-documented processes for human oversight, including clear pathways to human support and defined expectations for review of user-facing AI outputs.

Vendors must match risk to controls, and plan sponsors should ensure that higher risk, user-facing generative AI tools have tighter oversight, stronger guardrails and clear escalation to clinical teams. Plan sponsors should require vendors to include human backstops in AI solutions, with clear pathways to humans and defined expectations for reviewing participant-facing outputs.

Final Thoughts

As AI capabilities continue to grow, so does the responsibility to make sure these tools are used carefully and safely, with proper plans in place to reduce risk. For plan sponsors, this means feeling confident when asking for clear, ongoing information about how AI is being used by mental health solutions. As fiduciaries, plan sponsors could be liable if AI tools are implemented without appropriate oversight, due diligence and participant protections. By setting clear expectations with vendors about AI oversight, disclosure, data handling and accountability, plan sponsors can help balance innovation in care delivery while keeping plan participants safe. Self-directed educational materials and chatbots cannot truly replicate the therapeutic alliance formed with mental health professionals. It is crucial for mental health services to be easily accessible and available to participants to ensure any potential mental health issues are addressed before they deteriorate.

AUTHORS

Sandy Arena, RN, is a clinical consulting analyst on Segal's national health consulting and analytics team. Before joining Segal, she supported AI-powered prior authorization processes and conducted clinical documentation review at a health tech start-up. Arena holds experience in health policy, nursing research, clinical operations and direct patient care, with expertise in community and public health. At Segal, she supports vendor procurement for on-site and near-site clinics, point solutions, and utilization management and case-management programs. She holds a master's degree in public health. Arena can be reached at sarena@segalco.com.

Sarah Gunderson, RN-BC, NI-BC, is vice president of clinical consulting in the health practice at Segal. Her experience in nursing includes behavioral health, health care informatics and health analytics. Gunderson focuses on clinical audits, including behavioral health and the clinical portion of mental health parity audits. She also works closely with Segal's Health Analysis of Plan Experience (SHAPE) team in enhancing clinical reports and data analytic capabilities. She holds a master's degree in health informatics. Gunderson can be reached at sgunderson@segalco.com.

Ultimately, plan sponsors may achieve the desired balance of AI with compassionate clinicians. Mental health benefits will still be lacking without careful deployment that appeals to participants' interests while dispelling stigma that may prevent use. The true success measure of any mental health program should not be the number of AI-powered solutions, but whether those solutions help participants feel supported, understood and able to access care when they need it. BQ

Authors' note: The authors acknowledge the contributions of their colleagues Kaulen Pickell, vice president of communications at Segal, and Michael Stoyanovich, vice president of administration and technology consulting and senior consultant at Segal, for their peer review of this article and the AI expertise they shared.

Endnotes

1. Windsor Johnston, "With Therapy Hard to Get, People Lean on AI for Mental Health. What Are the Risks?" NPR. September 30, 2025. www.npr.org/sections/shots-health-news/2025/09/30/nx-s1-5557278/ai-artificial-intelligence-mental-health-therapy-chatgpt-openai.

2. Kaan Ozcan, "Using AI for Advice or Other Personal Reasons is Linked to Depression and Anxiety." NBC News. January 20, 2026. www.nbcnews.com/health/mental-health/ai-chatbots-personal-support-linked-depression-anxiety-study-rcna255036.

3. "One in Eight Adolescents and Young Adults Use AI Chatbots for Mental Health Advice." RAND Health. Accessed February 25, 2026. www.rand.org/news/press/2025/11/one-in-eight-adolescents-and-young-adults-use-ai-chatbots.html.

4. Sarah Wells, "Exploring the Dangers of AI in Mental Health Care." Stanford Graduate School of Education. June 11, 2025. https://ed.stanford.edu/news/exploring-dangers-ai-mental-health-care.

5. Zara Abrams, "Using Generic AI Chatbots for Mental Health Support: A Dangerous Trend." American Psychological Association. March 12, 2025. www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists.

6. Kate Wells, "An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. June 9, 2023. www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea.

7. "Experts Caution Against Using AI Chatbots for Emotional Support." Teachers College of Columbia University. Accessed February 19, 2026. www.tc.columbia.edu/articles/2025/december/experts-caution-against-using-ai-chatbots-for-emotional-support.

8. Ben Tarnoff, "Weizenbaum's Nightmares: How the Inventor of the First Chatbot Turned Against AI." *The Guardian*. July 25, 2023. www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai.

9. Mario Aguilar, "Why Woebot, a pioneering therapy chatbot, shut down." STAT News. July 2, 2025. www.statnews.com/2025/07/02/woebot-therapy-chatbot-shuts-down-founder-says-ai-moving-faster-than-regulators.

10. Executive Summary for the Digital Health Advisory Committee Meeting: Generative Artificial Intelligence Enabled Digital Mental Health Medical Devices. U.S. Food & Drug Administration. November 6, 2025. www.fda.gov/media/189391/download.

11. Zara Abrams, "Using Generic AI Chatbots for Mental Health Support: A Dangerous Trend." American Psychological Association. March 12, 2025. www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists.

12. Michael V. Heinz, M.D., Daniel M. Mackin, Ph.D., Brianna M. Trudeau, B.A., Sukanya Bhattacharya, B.A., Yinzhou Wang, M.S., Haley A. Banta, Abi D. Jewett, B.A., Abigail J. Salzhauer, B.A., Tess Z. Griffin, Ph.D. and Nicholas C. Jacobson, Ph.D. "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment." *New England Journal of Medicine*, March 27, 2025. https://ai.nejm.org/doi/10.1056/AIoa2400802.

13. Kameryn Griesser, "Your AI therapist might be illegal soon. Here's why." CNN Health. August 27, 2025. www.cnn.com/2025/08/27/health/ai-therapy-laws-state-regulation-wellness.

International Society of Certified Employee Benefit Specialists

Reprinted from the Third Quarter 2026 issue of *BENEFITS QUARTERLY*, published by the International Society of Certified Employee Benefit Specialists. With the exception of official Society announcements, the opinions given in articles are those of the authors. The International Society of Certified Employee Benefit Specialists disclaims responsibility for views expressed and statements made in articles published. Subscription information can be found at iscebs.org.

pdf/826