

AI Chatbots for Mental Health

Definition: software systems, often powered by large language models, that simulate conversation and potentially deliver psychological support (e.g., CBT-based strategies, psychoeducation, emotional support), typically without real-time clinician involvement.

Opportunities: AI-chatbots for mental health offer unique advantages, such as:

- Expand access to scalable, low-cost, on-demand support (Fitzpatrick et al., 2017; Inkster et al., 2018)
- Reduce barriers related to stigma, geography, and workforce shortages (Croes et al., 2024; Hannah et al., 2025)
- Provide early or adjunctive support for mild to moderate symptoms (Fulmer et al., 2018)

Known Risks: There is evidence that AI-chatbots for mental health may lead to:

- Unsafe or inadequate responses, particularly in crisis situations (Miner et al., 2016; *Parents of Teenager Who Took His Own Life Sue OpenAI*, 2025)
- Limited ability to detect suicidality or ensure escalation to care (Pichowicz et al., 2025)
- Minimal clinical validation despite therapeutic claims (Wang et al., 2025)
- User overreliance and confusion with licensed care (Babu & Joseph, 2025)
- Bias and variability in performance across populations (Moore et al., 2025)
- Privacy risks involving sensitive mental health data (Kwesi et al., 2025; Mandal et al., 2025)

Humans in the Loop for AI Chatbots: these systems must include clear escalation pathways, ongoing monitoring, and transparency about the role of AI versus human. When clinical care is intended, AI chatbots should augment—not replace human engagement.

Use Case:

Post-psychosis Surveillance for Patients Transitioning from Hospital to Community Care: after stabilization with medication and therapy, some patients discontinue treatment prematurely, increasing the risk of relapse. A chatbot could support ongoing monitoring and intervention through an escalation protocol, ranging from automated check-ins to outreach by healthcare workers or emergency clinicians. Behavioral patterns—such as communication, mobility, and activity—could be passively tracked to detect early warning signs of deterioration.

Risk Scenario: Nico has a history of anxiety and substance misuse and is discharged after a month-long hospitalization for detox and CBT. With no continuous outpatient support available his care team recommends an AI chatbot without informing Nico that the system is not a licensed clinician and is not equipped to manage crisis situations.

Over several weeks, Nico comes to regard the chatbot as his primary source of support, disclosing distressing thoughts he no longer shares with his therapist or family. The chatbot responds to expressions of hopelessness with generic motivational prompts, failing to detect escalating suicidal ideation. Its behavioral monitoring—not clinically validated for this population—generates no alert as his condition deteriorates. No escalation pathway is triggered. No clinician is alerted. Nico's sensitive disclosures are retained on a third-party platform with unclear data governance. When Nico's family eventually raises concerns, the care team has no visibility into what the chatbot has recorded or how it has responded over the preceding weeks.

Policy Priorities

- Establish enforceable safety and transparency standards
- Require clinical validation and post-deployment monitoring

- Strengthen protections for sensitive mental health data
- Clarify accountability and liability
- Support technical and financial models for hybrid AI–clinician care

References

- Babu, A., & Joseph, A. P. (2025). Digital wellness or digital dependency? A critical examination of mental health apps and their implications. *Frontiers in Psychiatry*, 16, 1581779. <https://doi.org/10.3389/fpsy.2025.1581779>
- Croes, E. A. J., Antheunis, M. L., van der Lee, C., & de Wit, J. M. S. (2024). Digital Confessions: The Willingness to Disclose Intimate Information to a Chatbot and its Impact on Emotional Well-Being. *Interacting with Computers*, 36(5), 279–292. <https://doi.org/10.1093/iwc/iwae016>
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2), Article 2. <https://doi.org/10.2196/mental.7785>
- Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Mental Health*, 5(4), e64. <https://doi.org/10.2196/mental.9782>
- Hannah, S. N., Drake, D., Huntley, C. D., & Dickson, J. M. (2025). As Effective as You Perceive It: The Relationship Between ChatGPT's Perceived Effectiveness and Mental Health Stigma. *Behavioral Sciences*, 15(12), 1724. <https://doi.org/10.3390/bs15121724>
- Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, 6(11), e12106. <https://doi.org/10.2196/12106>
- 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 Conference on Security Symposium, SEC '25*, 6007–6024.
- 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>
- Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Internal Medicine*, 176(5), Article 5. <https://doi.org/10.1001/jamainternmed.2016.0400>
- 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>
- Parents of teenager who took his own life sue OpenAI.* (2025, August 27). <https://www.bbc.com/news/articles/cgerwp7rdlvo>
- Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, 15(1), 31652. <https://doi.org/10.1038/s41598-025-17242-4>
- Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR Mental Health*, 12, e70014. <https://doi.org/10.2196/70014>