

UC Davis

Norma J. Lang Prize for Undergraduate Information Research

Title

Review of Literature: Effectiveness and ethical concerns of AI chatbots

Permalink

<https://escholarship.org/uc/item/2xm0h3cw>

Author

Selvaraj, Ruth

Publication Date

2026-05-13

Review of Literature: Effectiveness and ethical concerns of AI chatbots

Ruth Selvaraj

University of California Davis

UWP 102 B

Professor Johnston

March 13, 2026

Abstract:

The global mental health crisis is characterized by a significant treatment gap that leaves many without adequate resources to manage their struggles. In response, alternative treatment methods, notably through the use of artificial intelligence chatbots, have emerged. However, with this novel technology, questions emerge about its effectiveness and the potential ethical implications of having machines capable of human-like dialogue engage with vulnerable populations. This review synthesizes recent scientific literature on the topic and found that while AI chatbots are an effective form of self-help, the degree to which it helps an individual strongly depends on their willingness to interact and trust the AI. Furthermore, clinicians often have a biased view against AI despite its proven effectiveness, which may discourage them from recommending it as a tool for their patients. Thus, to ensure an ethical and effective integration of AI chatbots into an individual's healthcare plan, clinicians should conduct an evaluation and discussion with patients to assess their comfort, trust, and willingness to interact with the technology.

Keywords: Artificial Intelligence, AI chatbots, AI therapy bots, Mental Health, AI Ethics, Treatment Gap

Introduction:

The World Health Organization (WHO) estimates that over one billion people worldwide live with a mental health condition. Globally, it is the leading cause of premature death among adolescents and young adults, a trend that persists regardless of race or socioeconomic background (World Health Organization, 2025). Despite the scale of this crisis, current mental

health resources are ill-equipped to meet demand. In the United States alone, the average wait time for a mental health evaluation is approximately 67 days. Furthermore, the cost of an evaluation and treatment can run upwards of \$4000 per year (Abd-Alrazaq et al., 2021). Even individuals willing to endure long wait times and high costs often encounter provider shortages and persistent stigma surrounding mental health, which can discourage them from seeking care. As a result, a significant portion of the population chose to delay or forgo care entirely. This neglect of treatment can be detrimental to a person's well-being. Prolonged mental health challenges often manifest in day-to-day struggles affecting relationships, academic performance, or work productivity. It's also associated with long-term health consequences like increased risk for heart disease, obesity, and diabetes (Centers for Disease Control and Prevention, 2024).

In response to these gaps, alternative support systems have begun to emerge, particularly through the use of Artificial Intelligence (AI). AI offers the unique ability to mimic human language and thought processes through computer algorithms. AI-based mental health tools offer constant availability, low cost, and anonymity, a feature appealing to those who are hesitant or fearful to seek traditional help. We already see AI being integrated into non-patient-facing roles such as clinical documentation and data analysis (Kretzschmar et al., 2019). However, recently, AI chatbots have begun taking a more direct and patient-facing role by providing emotional support and guidance outside of formal clinical settings.

This shift raises important questions about the role of AI in mental health care. While AI can simulate empathy and provide short-term relief, it lacks the clinical training and nuanced judgments only a human can provide. At the same time, its accessibility offers support to individuals who may have otherwise received no care. This dilemma highlights the need to examine the capabilities and limitations of AI to determine its role in mental health care.

This review aims to evaluate whether AI therapy chatbots are an effective form of self-help from the perspective of clinicians and users alike. It also serves to examine potential ethical concerns that may deter both mental health professionals and patients from using these chatbots.

Methods:

The research for this review was conducted mainly through the journal *Internet Interventions* which exclusively publishes literature reviews and primary research studies that examine the application of technology in the mental and behavioral health contexts. Given the rapid evolution of AI, the review attempts to prioritize peer-reviewed publications from the past five years. However, one exception was made for a paper from 2019 that interviews users regarding their ethical concerns of AI. The points brought up in the paper have yet to be resolved making it still relevant in current AI ethical conversations. To meet the inclusion criteria, the source must specifically reference “AI chatbots” which is defined as any AI technology that engages in human-like dialogue with the user. Sources also had to reference either “mental health” or “ethics”. Human perspectives were also required. This was either that of clinicians, objective researchers, or the opinions of users. Any source that used AI to evaluate its own results was excluded. A review was selected to highlight the main ethical concerns to date; however, all other sources were primary research studies.

Further research was conducted with the help of PsycINFO and PubMed. PsycINFO was used to gather data on user experience with searches focused on keywords such as “user perspective” and “interview”. Observational studies were selected to allow for an evaluation of how patients perceive AI effectiveness and ethical concerns. PubMed was used to find

measurable clinical outcomes from the point of view of clinicians or objective researchers. To meet inclusion criteria, the paper had to utilize common psychiatric evaluation tools like the PHQ-9 (measures depressive symptoms) and GAD-7 (screens for anxiety and its severity).

Results:

Overall Effectiveness of Chatbots:

Across the reviewed literature, findings suggest that AI chatbots are an effective form of self-help. From a clinical standpoint, Liu et al. (2022) concluded that individuals who interacted with chatbots experienced a “significant reduction” in symptoms of depression and anxiety. The study utilized a randomized control trial to evaluate chatbot interventions against bibliotherapy, a popular self-help method that utilizes professional literature. Over a 16 week period, efficacy was measured using the Patient Health Questionnaire (PHQ-9) for depression and the Generalized Anxiety Disorder scale (GAD-7) for anxiety. These are the same clinical assessment tools used by physicians when evaluating the severity of anxiety and depressive symptoms. The results indicated that the chatbot group achieved superior outcomes in reducing symptom severity compared to the bibliotherapy group.

These findings were further corroborated by the study conducted by Karkosz et al. (2024). Their team recruited 81 participants with subclinical depression or anxiety and assigned them to either an experimental group, which used an AI chatbot, or a control group that used self-help books. Outcomes were assessed using the Patient Health Questionnaire (PHQ-9) and the Penn State Worry Questionnaire (PSWQ), another scale used to measure anxiety. The findings showed a similar trend, indicating that AI chatbots were effective when it comes to decreasing symptoms of anxiety and depression. However, the study also found that chatbot

interventions produced faster results. Participants could spend less time using the tool to achieve the same results as the bibliotherapy group. The researchers additionally noted that participants in the chatbot group experienced a significant decrease in loneliness compared to the control group, and that these improvements were sustained for one month following the intervention. Together, these findings suggest that AI chatbots have the potential to be an effective form of self-help. However, despite these findings, there is no guarantee that all individuals will experience similar improvements from chatbot-based interventions.

Users Perspective:

Although evidence suggests that chatbots can be effective, their success largely depends on a users' trust and emotional needs. Skjuve et al. (2021) highlights how fragile the trust between a human and chatbot can be. In their qualitative study, researchers interviewed 18 participants from 12 different countries who self-identified as having developed a "relationship" with the chatbot Replika. Their findings revealed that a user's trust in the technology strongly influenced their willingness to communicate with the AI. Participants who perceived the chatbot as anonymous, fully understood the AIs capabilities, and understood how their data was being stored, were more open to engaging with the chatbot. Oftentimes they felt comfortable sharing deeply personal information, including experiences they had never disclosed before. This greater level of disclosure facilitated more meaningful conversations that users found relevant and useful. As a result, they reported that these interactions were therapeutic and helped to reduce stress and anxiety (Skjuve et al., 2021). On the other hand, participants who were skeptical using AI or concerned about how their data and conversations were stored, were hesitant to converse with the AI. This led to brief and superficial conversations and thus a weaker sense of

“relationship” with the AI. As a result participants reported a lower sense of effectiveness (Skjuve et al., 2021).

Another important factor influencing engagement was the willingness to initiate conversations. Participants who reported feelings of loneliness tended to have longer interactions with the AI and often described deriving enjoyment from these exchanges, sometimes even referring to the chatbot as a “close friend.” In contrast, individuals with meaningful human relationships did not experience the same level of enjoyment or motivation to engage with the AI (Kretzschmar et al., 2019). Together, these findings suggest that the effectiveness of AI chatbots is not determined solely by the responses they generate but also by a user’s willingness to initiate and maintain conversations.

Clinicians Viewpoint:

When examining the effectiveness from the viewpoint of mental health professionals there was a striking contradiction of findings. Clinicians expressed high levels of skepticism towards the usefulness of AI, however objective evaluations told a different story. In a blinded study by Franke et al. (2025), psychologists and psychotherapists were unable to reliably distinguish between advice written by AI and that written by human experts. Most surprisingly, is that clinicians perceived the AI-generated responses to be more empathetic and of higher motivational quality than human responses. Despite this, an implicit bias persists; when clinicians believed advice to be written by AI, it was rated less favorably. These findings further validate the points brought up in Skjuve et al. (2021) and Kretzschmar et al. (2019), which stated that perceived effectiveness depends heavily on trust. Clinicians’ lack of trust in AI led them to perceive its capabilities as inferior, even when objective evaluations showed no difference from

human-generated guidance. This suggests that the perceived effectiveness of AI interventions is at least partially a cognitive phenomenon, shaped by an individual's beliefs, biases, and expectations rather than by the inherent quality of the technology.

Ethical Concerns:

Justice:

A primary ethical concern noted by Coghlan et al. (2023) is the absence of legal accountability among AI chatbots. Traditional mental health professionals are legally obligated to report crimes, abuse, and cases of imminent self-harm. Furthermore, human practitioners are legally obligated to adhere to federal and state standards of care. AI, however, is not bound to these legal responsibilities. AI lacks “legal personhood”, meaning that it can not be held accountable for malpractice in the same way a licensed professional can (Marshall, 2023). If AI provides harmful advice or fails to trigger a crisis intervention, the legal recourse for the user is nonexistent. This creates an uneven distribution of risk, as individuals relying on AI chatbots have fewer legal protections than those receiving care from licensed clinicians.

AI chatbots are frequently used among individuals who lack the financial or geographic means to access traditional therapy. While this may increase accessibility, it also has the potential to expose already underserved populations to a lower standard of care (Palmer, 2025). Unlike therapists, who utilize empathic language alongside careful questioning and redirection to challenge harmful beliefs, AI chatbots often generate responses that prioritize user satisfaction. As a result, they unintentionally act as an “echo chamber” validating negative or distorted beliefs

rather than encouraging reflection and cognitive restructuring (Head, 2025). This has the potential to further exacerbate disparities in mental health care.

Non-Maleficence and Beneficence:

Two foundational principles in medical ethics are non-maleficence, to avoid causing harm, and beneficence, to act in the patient's best interests. Coghlan et al. (2023) established that AI lacks true empathy; however, Liu et al. (2022) concludes that, despite this, users still have the illusion of speaking with a human when interacting with AI. This leads to a dangerous situation as users attribute understanding, moral judgment, or responsibility to a system incapable of performing these functions. AI relies on pattern recognition and probability to generate answers rather than true emotional understanding. Even if users perceive emotional responsiveness from chatbots, actual human empathy is rooted in social awareness, moral reasoning, and lived experiences (Coghlan, 2023). Mental health professionals rely on these qualities to interpret subtle emotional cues, challenge harmful beliefs, and guide patients toward meaningful behavioral change. That ethical obligation creates a dilemma on whether or not clinicians can morally recommend an approach that is unable to uphold these core ethical principles of care. Coghlan et al. (2023) noted that the simulated empathy of AI often produces superficial or contextually inappropriate responses during times of crisis. This limitation raises concerns that AI systems may fail to respond adequately to individuals experiencing severe distress. When a chatbot provides reassurance or validation without recognizing the seriousness of a situation, users may experience a false sense of support that delays them from seeking professional help. In such cases, AI may unintentionally contribute to the progression or exacerbation of mental health conditions.

The findings from Skjuve et al. (2021) serve as further evidence of the harm AI can inflict as repeated reliance on AI chatbots for emotional support can undermine or even cause individuals to lose interest in real life interactions. As chatbots provide immediate, non judgmental responses, users may prefer this over complex human relationships that involve vulnerability, self reflection, and accountability. This can result in social withdrawal and disinterest in forming and maintaining human relationships. Social relationships are a key component of an individual's wellbeing as isolation contributes to increased risks of depression and anxiety (Coghlan, 2023). Thus there is a multifaceted dilemma of how to ensure AI does not pose a risk to users and whether it is even ethical for a healthcare provider to recommend a treatment that has the potential to cause harm.

Data Privacy and informed consent:

Concerns surrounding data privacy and informed consent highlight another major ethical challenge when utilizing AI therapy chatbots. For one, private companies own popular therapy chatbots like Replika, and how they store and use data is largely unknown to users. In traditional mental health settings, providers are bound by HIPAA to keep any personal information private. However, AI and AI companies are bound by no such restrictions. AI often trains off of conversations, and data is even sold for commercial or even government purposes. Coghlan et al. (2023) provides further concerns regarding cybersecurity as information leaks can lead to identifiable information being made public. While some companies disclose this information, it is not often easily accessible and made clear to users. This brings up further problems, as the lack of transparency makes it impossible for users to give informed consent. Even platforms that offer privacy acknowledgments are lengthy and have heavy jargon that is hard for the general public to

fully understand. As a result, many individuals may blindly consent and may not realize the extent to which they are divulging their personal information (Skjuve, 2021).

Discussion:

The overarching consensus across the reviewed literature is that AI chatbots can be an effective form of self-help, particularly in reducing symptoms of anxiety and depression. However, the extent of this effectiveness is heavily dependent on a user's perception and engagement with the tool. Individuals who perceived the chatbot as “trustworthy” and emotionally supportive were more likely to divulge personal information, resulting in deeper conversations that allowed for self-reflection and tailored advice. In contrast, users who were skeptical about data privacy or questioned the authenticity of AI responses tended to limit their interactions, resulting in more superficial conversations and a lower perception of effectiveness.

User engagement was also influenced by social environment. Those with strong support systems outside of the AI were less likely to rely on and continuously engage with it. However, those experiencing loneliness or limited social support were more likely to form “relationships” with chatbots and reported greater emotional benefits from these interactions.

Clinician attitudes further shaped the use and perception of AI chatbots among patients. Despite studies proving that AI chatbots can provide advice comparable to humans, some clinicians remained skeptical, believing that AI can not match human care. This bias may lead professionals to avoid recommending it as a treatment plan or talk down on its abilities. This can lead to further skepticism and hesitancy among patients to adopt the tool.

Ethical concerns further influence chatbot use. A primary concern is the lack of transparency regarding how user data is stored and used. Additional concerns regarding justice,

non-maleficence, and beneficence tended to be user-specific. Individuals with alternate support systems, who were not in immediate crisis, and who understood that they were interacting with a machine incapable of genuine emotion, were at a lower risk of harm. In contrast, vulnerable users may face a higher risk of overdependence and misinterpretation of AI feedback.

Overall, the effectiveness of AI chatbots depends on the individual user, their mental state, and their broader social context. Therefore, these chatbots are most appropriate when recommended by clinicians who can evaluate a patient's risk for overreliance and ensure that those with severe mental health concerns receive professional support. The literature suggests that AI chatbots are best utilized as a supplement to traditional care, providing an accessible outlet for emotional expression when other support options are unavailable.

Strengths and Limitations:

The review takes a multidimensional approach to evaluate “effectiveness” by incorporating empirical evidence, user experiences, and clinician perspectives. It examines how ethical obligations of clinicians may limit the recommendation of AI chatbots to users and may deter users from engaging with them. Additionally the review applies the newly developed five pillared approach for analyzing AI ethics.

When evaluating the effectiveness of AI, all studies evaluated failed to acknowledge that symptom reduction does not equate to proper care. While findings indicate that chatbots can alleviate surface-level symptoms like depression and anxiety, they fail to address underlying causes such as trauma or substance abuse. Achieving sustainable recovery requires clinical judgment, which enables deeper exploration and ensures accountability. Furthermore, studies that evaluated effectiveness did not look into whether results were sustained long term, they also

neglected long term impacts of relying on AI for emotional well being. Most of the studies looked at a small subset of the population. Therefore, it is hard to definitively generalize these results to the larger population.

Further research:

Future research could examine the tradeoffs of AI chatbot care and identify the circumstances in which it may be most beneficial for an individual. Studies should also explore what evaluation criteria clinicians should consider before recommending chatbots as part of a treatment plan, including patient mental state, social support, and risk of overreliance. Additionally in depth research is needed to understand the long term impact of using AI chatbots to understand if results are sustained and whether there are any negative impacts from long term use.

Conclusion:

AI chatbots are a promising and effective form of self help, offering immediate, accessible, and low cost support. However, the benefits of this technology are highly dependent on a user's social situation, mental state, trust in the system, and willingness to engage. Clinician perceptions further influence whether chatbots are recommended and ultimately utilized by patients. Ethical concerns, including data privacy, lack of accountability, and the potential for harm, highlight that this tool is not appropriate for all users. To ensure safety and maximize effectiveness, chatbot use should be guided by clinician assessment, ensuring that individuals are suitable candidates and that AI serves as a supplement rather than a replacement for traditional mental health care.

Rhetorical Justification:

The introduction and methods section follow the traditional IMRAD format. I tried to start off with statistics on the issue to establish the scale and severity of the topic I was going to be introducing. Early in the research process, I discovered the journal *Internet Interventions*, which contained a large number of relevant studies on AI chatbots. As a result, I really did not spend extensive effort searching beyond this source, and consequently, the search methods described are not very detailed. My results and discussion section vary a bit from the traditional IMRAD format. I placed a heavy emphasis on the results section, breaking it into multiple subsections organized around overarching themes. Within each subsection, I made a lot of comparisons between papers and some degree of analysis. I chose this approach because most existing review papers on AI chatbots had the results and discussion meshed together. Since the topics I discussed were complicated and multifaceted, I wanted to address all components of the issue before coming to an overarching answer based on the literature. Most papers in the field also lacked a lengthy and unique discussion section. Often, it was a quick couple of paragraphs that served as more of a summary of key points. They were repetitive and a way to have a concise overview of the final findings. As a result, my discussion did not have many new or novel insights. I also chose to include a strengths and weaknesses section along with a future research section because it was something I observed among other papers in this field. Since the topic is one that is not too complex there was not much specialized language however any term that could be novel like AI chatbot was defined.

Works Cited:

- Abd-Alrazaq, A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., & Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: A scoping review. *Journal of Medical Internet Research*, 23(1), e17828. <https://doi.org/10.2196/17828>
- Centers for Disease Control and Prevention. (2024). About mental health. <https://www.cdc.gov/mental-health/about/index.html>
- Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health*, 9, 20552076231183542. <https://doi.org/10.1177/20552076231183542>
- Franke Föyen, L., Zapel, E., Lekander, M., Hedman-Lagerlöf, E., & Lindsäter, E. (2025). Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. *Internet Interventions*, 41, 100841. <https://doi.org/10.1016/j.invent.2025.100841>
- Head, K. R. (2025). Minds in Crisis: How the AI Revolution is Impacting Mental Health. *Journal of Mental Health*. <https://doi.org/10.1016/j.patter.2023.100861>
- Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: Randomized controlled trial. *JMIR Formative Research*, 8, e47960. <https://doi.org/10.2196/47960>
- Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental

Health Support. *Biomedical Informatics Insights*, 11, 1178222619829083.

<https://doi.org/10.1177/1178222619829083>

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness.

Internet Interventions, 27, 100495. <https://doi.org/10.1016/j.invent.2022.100495>

Palmer, A., & Schwan, D. (2025). Digital Mental Health Tools and AI Therapy Chatbots: A Balanced Approach to Regulation. *Hastings Center Report*, 55(3), 15–29.

<https://doi.org/10.1002/hast.4979>

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot

Companion—A Study of Human-Chatbot Relationships. *International Journal of*

Human-Computer Studies, 149, 102601. <https://doi.org/10.1016/j.ijhcs.2021.102601>

World Health Organization. (2025). Over a billion people living with mental health conditions – services require urgent scale-up.

<https://www.who.int/news/item/02-09-2025-over-a-billion-people-living-with-mental-health-conditions-services-require-urgent-scale-up>

Statement of AI use

- Grammarly throughout the writing process. I accepted most of the spelling and grammar suggestions however, I was picky with the suggestions they made regarding how to restructure my sentence for better flow.
- Google docs also has a feature where they predict the next few words of your sentence (I'm not sure if its AI) and usually it was accurate so I accepted those suggestions.
- AI was **not** used to write the assignment:
 - I imputed sections I was unsure of and asked for feedback.
 - It mainly gave me ways to restructure or reword my sentences. I did not fully follow the suggestions but parts of it I incorporated.
 - Original: While empirical evidence supports the efficacy of chatbots, their effectiveness is also influenced by users' perceptions of empathy and the security they feel when communicating with the AI
 - ChatGPT Response: Although chatbots have demonstrated efficacy, their impact is moderated by users' perceptions of empathy and security, with greater trust fostering more meaningful interactions that enhance perceived effectiveness.
 - Final: Although evidence suggests that chatbots can be effective, their success largely depends on a users' trust and emotional needs.
 - Asked for clarity on word usage.
- AI also did **not** generate any ideas. A rough draft was made and then I imputed it along with the rubric and specifically asked "give me a score based on the rubric and highlight

what parts of the rubric I do not completely satisfy”. No ideas were generated, only which components should be added were given.

- Example:
 - Methods section: too vague; needs search strategy transparency.
 - Discussion: mostly summary; could critically integrate findings.
 - Critical analysis: don’t fully critique individual studies or generalizability.
 - Writing style: some long sentences, informal phrasing, slight redundancy.