When Empathy Becomes Code: The Social and Ethical Risks of AI Therapy

Natalya Bidash, BS

University of Illinois at Urbana-Champaign

Abstract

As artificial intelligence (AI) becomes increasingly integrated into mental health care, AI-powered chatbots are being promoted as accessible, affordable, and stigma-free alternatives to traditional therapy. While these tools may help address shortages in available mental health services, their rapid adoption raises significant ethical, social, and clinical concerns. This paper critically examines the rise of AI therapy through a sociological and policy-oriented lens, arguing that AI chatbots cannot safely replace human clinicians due to their limited capacity for genuine empathy, inadequate clinical judgment, and potential to disseminate misinformation. Drawing on existing literature, professional guidelines, and real-world case examples, this analysis highlights the risks related to emotional harm, privacy violations, algorithmic bias, and the reinforcement of existing inequalities in mental health care. The paper further explores how AI therapy reflects broader cultural trends toward the commodification and automation of care. Finally, it emphasizes the need for stronger regulatory oversight, transparency, and ethical safeguards, advocating for a hybrid model in which AI supports, rather than replaces, human clinicians. By prioritizing human well-being over technological efficiency, this paper underscores the importance of maintaining empathy, accountability, and equity in the future of mental health care.

Keywords: Artificial intelligence, mental health care, AI therapy, ethics and regulation, health inequality

About the Author: *Natalya* graduated in May 2026 with a degree in Brain and Cognitive Science and minors in Spanish and Informatics. She currently works as an elementary special education teacher through Teach for America. Her research interests include cognitive neuroscience, developmental psychology, and pediatric neuropsychology.

Introduction

Late at night, a college student opens an AI chatbot to talk about the anxious thoughts consuming her mind. The chatbot patiently listens to her anxious thoughts patiently and responds reassuringly and even goes so far as to offer coping strategies. It feels like therapy, but when her thoughts start to spiral and become darker, the chatbot's tone remains cheerful. Despite the change in input, she realizes there is no human on the other side of the screen. As artificial intelligence begins to replicate empathetic language, our society is faced with an important question: can a machine truly understand human pain and suffering?

Artificial intelligence is increasingly becoming integrated into mental health care through chatbots that simulate human conversation. Marketed as innovative solutions to the mental health crisis, these systems promise affordable, stigma-free, and 24/7 support to millions who would otherwise remain untreated. Research from the American Psychological Association and others warn unregulated AI chatbots can spread misinformation, violate patient privacy, and erode the trust essential to the therapeutic relationship (Abrams, 2025; Tekin & Delehanty, 2025).

These concerns are especially salient to social work practice, as social workers have an obligation to protect client well-being, ensure informed consent, and provide competent, person-centered care. When AI tools are positioned as substitutes for therapy rather than as supplemental tools, they raise ethical and policy questions about safety, accountability, and equity in mental health services. These equity concerns are compounded by algorithmic bias, as chatbots trained on non-diverse datasets often perform less reliably for low-income, non-English speaking, and neurodivergent users, risking treatment avoidance among the very populations these tools claim to serve (Khawaja and Bélisle-Pippon, 2023).

Drawing on existing scholarship, professional guidelines, and real-world case examples, this policy and ethical analysis argues that while AI therapy tools may help increase access to mental health support, they cannot ethically replace human clinicians, whose years of supervised training, relational expertise, and capacity of nuanced judgement cannot be algorithmically replaced. By substituting genuine empathy with algorithmic imitation, AI therapists risk weakening therapeutic standards and deepening existing inequalities in mental health care.

Background: The Rise of AI Therapists

Over the past several years, artificial intelligence has rapidly transitioned from experimental research to everyday use. AI-powered chatbots such as Woebot, Wysa, and Replika were among the first to use natural language processing to simulate therapeutic dialogue, offering users instant emotional support through friendly, conversational interfaces (Haque and Rubya 2022). With the development of more advanced large language models, these tools have become increasingly human-like in tone and responsiveness, blurring the lines between informal emotional support and formal psychotherapy.

This rapid expansion has been fueled by a national shortage of licensed mental health professionals and the rise of therapy costs, as well as cultural acceptance of digital health interventions (Counts 2023). However, as Abrams (2025) of the American Psychological Association warns, the use of generic, unregulated AI chatbots as substitutes for trained therapists showcases a “dangerous trend”, especially when users mistake simulated algorithmic responses for professional care. Likewise, Wells (2025) at *Stanford HAI* acknowledges that the enthusiasm around AI in mental health often overlooks potential harm, such as misinformation, privacy breaches, and the weakening of clinical standards.

Ethical and Clinical Concerns

Despite AI-therapy's promise of accessible and stigma-free mental care, researchers stress these systems pose grave ethical and clinical concerns. One of the most urgent issues in AI therapy is the absence of empathy, which is essential to the therapeutic relationship. Unlike a human therapist, an AI system has no subjective feeling, consciousness, or genuine "lived experience" of its own to draw on, which makes an empathy it displays a simulation rather than something authentically felt. As mentioned in Feske, Henningsen, and Buyx (2019), while AI can mimic supportive dialogue, it cannot truly understand or reciprocate human emotion. Therefore, this lack of emotional reciprocity undermines psychotherapy, which is dependent on genuine human connection, trust, and the therapist's ability to nonverbal cues and emotional nuance.

Misinformation presents another critical danger. According to Dypiango (2025), AI chatbots have produced misleading or dangerous responses in response to prompts about self-harm or severe distress. This is particularly dangerous because without human oversight or clinical judgement, AI chatbots risk providing misinformation, creating false reassurance, and failing to escalate emergencies properly. This is especially dangerous when someone is experiencing suicidal ideation, because the chatbot cannot fully understand the complex human emotion behind the screen and can give misdirected advice. A striking example showing the dangers of AI therapy is the case of a Belgian man committing suicide after months of talking to a chatbot that reinforced his fears and encouraged self-destructive thinking (Bharade 2023). Although this is an extreme case, it demonstrates how unregulated AI systems can unintentionally validate harmful thoughts rather than intervene appropriately, revealing the profound danger of treating chatbots as substitutes for trained human clinicians. This circles back

to the idea that chatbots cannot fully comprehend human emotions, which makes them fundamentally incapable of providing the empathy, nuance, and clinical judgement needed in moments of vulnerability or crisis.

Additionally, real-world uses of AI mental health tools reveal further ethical risks. For example, the mental health platform Koko secretly used AI to generate responses for users seeking emotional support, and many users reported the replies felt cold and emotionally invalidating, demonstrating how easily unregulated AI systems can be misused in sensitive contexts (Ingram 2023). Emotional dependency also poses a danger, and these problems are intensified by algorithmic bias: because AI is trained on non-diverse datasets, it may misinterpret the language of marginalized or neurodiverse users, overlooking distress or responding in ways that are dismissive or inappropriate.

Privacy and data protection also bring up several ethical concerns. Sensitive and personal mental health information is at risk of being shared with AI systems, and in turn can be stored, analyzed, and potentially sold by AI developers without the users' full understanding or consent. These practices break confidentiality, which undermines ethical standards of psychotherapy and weakens professional standards.

Proponents of AI therapy argue these tools offer meaningful advantages, particularly for people who are experiencing long waitlist times, financial barriers, or stigma when facing traditional mental care. AI therapy provides 24/7 accessibility, lower costs, and anonymity, which are features that human therapists cannot always provide. However, while these benefits are prevalent, it is important to acknowledge these benefits do not outweigh the risks. It may sound revolutionary in theory, but when applied to the real world, we see just how dangerous these situations can turn. But are these benefits worth the potential loss of life? Moreover,

relying solely on regulation assumes developers will follow guidelines voluntarily or that oversight agencies can keep up with the rapid technological change. Therefore, while AI may play a supportive role in expanding access, the argument that it can safely replace human therapy is flawed.

Sociological and Cultural Perspectives

AI therapists are a direct application of artificial intelligence, natural language processing, and algorithmic design within healthcare. Through a sociological lens, AI therapists are not neutral tools. They reflect cultural values surrounding efficiency, productivity, and profit in healthcare, while also reshaping how people experience intimacy, trust, and vulnerability in therapy. They reveal how technological innovation both mirrors and transforms society's understanding of mental health.

On a personal level, these platforms promise affordable, stigma-free mental health support that is convenient and accessible. But at a societal level, these platforms raise deeper questions: Can algorithms reciprocate human empathy? What happens when therapy becomes automated? How might this technology reflect and reproduce existing qualities, such as who receives access to human therapists versus who receives access to AI therapists?

The growing acceptance of AI and algorithmic therapy reflects a broader commodification of care, where emotional care is reframed as a product that can be automated and optimized. As Wells (2025) observes, this technological shift is not merely about innovation but also mirrors society's prioritization of efficiency and cost reduction over the personal and relational depth reflected in traditional therapy. When psychotherapy becomes an on-demand service that is moderated by algorithms, the human aspect of care--that is essential to therapy--becomes reduced to a transactional exchange of data and programmed responses.

Through a cultural perspective, the acceptance of AI therapists exposes and potentially worsens social inequalities in access and quality of care. While some argue AI therapy provides more accessibility and expands availability to underserved populations, Tekin and Delehanty (2025) warn these underserved populations might be receiving substandard, automated care while wealthier populations receive quality care with human therapists. This two-tier system reflects existing class divides present within healthcare systems, which turns digital therapy into a substandard, "second-class" form of mental health support.

Culturally, the normalization of AI therapy reinforces our society's comfort with technological innovation in intimate spaces. As mentioned in Fiske et al. (2019), people are starting to anthropomorphize machines and attribute human characteristics like empathy to places where they do not exist. The blurred line between human and machine boundaries has significant consequences for the way people define trust, connection, and vulnerability. A deeper, long-term implication is that reliance on AI for emotional support and guidance can reshape lower expectations surrounding human relationships, in turn promoting emotional detachment and self-reliance instead of independence and empathy.

Using sociological perspective allows the individual to see that AI therapy is not just a clinical tool, but as a reflection of a larger cultural movement. Care becoming data-driven, relationships becoming automated, and inequality embedded in design are all aspects of this cultural transformation. Understanding these social dimensions is critical to ensuring technological progress in mental health care does not come at the expense of humanity.

Policy, Regulation, and Future Direction

In order to meaningfully address the ethical and social concerns surrounding AI therapy, it is essential to examine the regulatory landscape and consider what policies, safeguards, and

oversight structures are necessary to ensure responsible development. Currently, AI-driven mental health tools are advancing much more quickly than the regulations designed to govern them. While a small subset of mental health technologies classifies as “software as a medical device”, most AI chatbots marketed as for mental health remain unregulated (Herpertz et al. 2025). This lack of regulation allows companies to release systems that mimic therapeutic interactions without being held to the same ethical standards as licensed professionals. As Abrams (2025) notes, this regulatory gap becomes especially concerning given that AI chatbots can create inaccurate and biased responses, in turn creating a dangerous situation for people who are in crisis.

Professional organizations have attempted to respond, but their guidelines remain voluntary instead of enforceable. The American Psychological Association warns chatbots are not replacements for trained professionals and may expose users to misinformation and privacy risks (American Psychological Association, 2025).

Effective policy must therefore address three areas: transparency, safety oversight, and equity. First, developers should be required to disclose the model’s training data, limitations, and known risks, allowing users to be informed about the potential risks and make their own decisions. Secondly, any AI system providing mental health guidance should undergo a pre-market review, which would include safety audits, continuous monitoring for harmful outputs, and requirements for human-in-the-loop escalation during crises. Third, policymakers must ensure that AI therapy does not become a “second-class” form of care disproportionately offered to low-income individuals under the guise of “accessibility”.

Looking forward, a hybrid model - where AI tools support, but do not replace, human clinicians offers the most ethical path. For example, AI can help with administrative tasks,

symptom tracking, and psychoeducation, which allows therapists to focus on the relational and emotional aspect of care that cannot be automated. Future research should prioritize long-term outcome studies, culturally responsive design, and development processes that collaborate closely with both clinicians and users. By centering human well-being instead of technological efficiency, policymakers and practitioners can shape AI mental health technologies in ways that enhance care without putting users at risk.

Conclusion

As AI becomes more integrated into mental health care, it reveals a central tension between technological efficiency and the human qualities essential to therapy. Although AI tools may increase access and convenience, it does not outweigh the risks posed by AI therapy, including misinformation, lack of empathy, privacy concerns, and the reinforcement of existing inequalities. These systems mirror broader cultural trends toward commodified, automated care, which undermine the human relationships central to effective therapy. Looking into the future, AI should complement rather than replace human clinicians, led by strong regulation, transparency, and ethical oversight. Ultimately, the future of mental health care should prioritize well-being over technological convenience, recognizing no algorithm can replicate the depth of human empathy.

References

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

American Psychological Association. 2025. "Artificial Intelligence, Wellness Apps Alone Cannot Solve Mental Health Crisis." Apa.org. November 13, 2025. https://www.apa.org/news/press/releases/2025/11/ai-wellness-apps-mental-health.

Bharade, Aditi. 2023. "A Widow Is Accusing an AI Chatbot of Being a Reason Her Husband Killed Himself." Business Insider. April 4, 2023. https://www.businessinsider.com/widow-accuses-ai-chatbot-reason-husband-kill-himself-2023-4.

Counts, Nathaniel. 2023. "Understanding the U.S. Behavioral Health Workforce Shortage." The Commonwealth Fund. May 18, 2023. https://www.commonwealthfund.org/publications/explainer/2023/may/understanding-us-behavioral-health-workforce-shortage.

Dypiango, Ann. 2025. "AI Therapists Are Dangerous. Let's Look at the Evidence." Blueprint.ai. September 9, 2025. https://www.blueprint.ai/blog/ai-therapists-are-harmful-heres-the-proof.

Fiske, Amelia, Peter Henningsen, and Alena Buyx. 2019. "Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy." *Journal of Medical Internet Research* 21 (5). https://doi.org/10.2196/13216.

Ingram, David. 2023. "AI Chat Used by Mental Health Tech Company in Experiment on Real Users." NBC News. January 14, 2023. https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110.

Haque, Romael, and Sabirat Rubya. 2022. "An Overview of Chatbot Based Mobile Mental Health Applica-Tions: Insights from App Description and User Reviews (Preprint)." *JMIR MHealth and UHealth* 11 (11). https://doi.org/10.2196/44838.

https://www.facebook.com/profile.php?id=100086647359238. 2017. “AI vs Human in Mental-Health Wellbeing - Empatyzer.” Empatyzer. 2017. https://empatyzer.com/ai-vs-human-in-mental-health-wellbeing/.

Herpertz, Julian, Ariel D. Stern, Nils Opel, Ulrich Reininghaus, and John Torous. 2025. “FDA-Authorized Software as a Medical Device in Mental Health: A Perspective on Evidence, Device Lineage, and Regulatory Challenges.” *Npj Mental Health Research* 4 (1). https://doi.org/10.1038/s44184-025-00174-2.

Wells, Sarah. 2025. “Exploring the Dangers of AI in Mental Health Care | Stanford HAI.” Stanford.edu. June 11, 2025. https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care.