Commentary

Big Data & Society
October–December: 1–5
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517251383884
journals.sagepub.com/home/bds

Artificial intelligence is mental: Evaluating the role of large-language models in supporting mental health and well-being

Briana Vecchione[1] and **Ranjit Singh**[1]

Abstract

As artificial intelligence-powered chatbots and large-language models become increasingly integrated into mental health support, they raise complex ethical, economic, and structural challenges—particularly related to dimensions of interaction, labor, and regulation. This commentary explores how users engage with artificial intelligence-driven chatbots for emotional well-being, the economic and professional implications for mental health providers, and the evolving regulatory challenges of artificial intelligence therapy. While chatbots offer the potential benefits of accessibility and personalized responses that resemble human conversation, it is crucial to question the limitations of their reliance on predictive modeling as it relates to their capacity for deep empathy and nuanced care. The growing adoption of artificial intelligence in mental health support threatens to reshape therapeutic labor by disrupting traditional therapy models while simultaneously reinforcing economic precarity among practitioners, and the absence of clear regulatory frameworks raises concerns about accountability, data privacy, and ethical use(s) of artificial intelligence-driven mental health tools. By analyzing these dimensions, this paper highlights the affordances and limits of artificial intelligence in mental health care and advocates for a critical approach to its development and governance.

Keywords

Artificial intelligence, large language models, mental health, therapy, chatbots, human–computer interaction

The use of large-language models (LLMs) like ChatGPT for therapy is rapidly expanding, sparking debate over their effectiveness and growing role in emotional care. As Figure 1 illustrates, some users report feeling genuinely understood by chatbots, enough to fully replace human therapists with artificial intelligence (AI) companions. For individuals seeking mental health support, the primary challenge often lies in emotional accounting: the process of making sense of, naming, and articulating what they feel and why. Traditionally, this has involved intimate conversations with loved ones, therapy sessions, or reflective practices like journaling. This commentary explores the emerging use of LLM-based chatbots as tools that promise to enhance mental health through guidance, emotional relief, psychological support, and self-help strategies.

Chatbots uniquely blend the privacy and comfort of personal diaries with the responsive engagement of conversation. They offer a judgment-free space that feels safe and private while also providing the interactivity crucial for self-reflection. Amid growing social media discussions about optimizing chatbot prompts for mental health, these systems are reshaping how people understand and practice mental wellness. This raises urgent questions: How are chatbots being integrated into people's daily routines? What forms of reliance, misinterpretation, or emotional attachment are emerging? And what are the social and ethical stakes as chatbots become more widely adopted for mental health support?

[1]AI on the Ground Program, Data & Society Research Institute, New York, NY, USA

Corresponding author:
Briana Vecchione, AI on the Ground Program, Data & Society Research Institute, New York, NY, USA.
Email: briana@datasociety.net

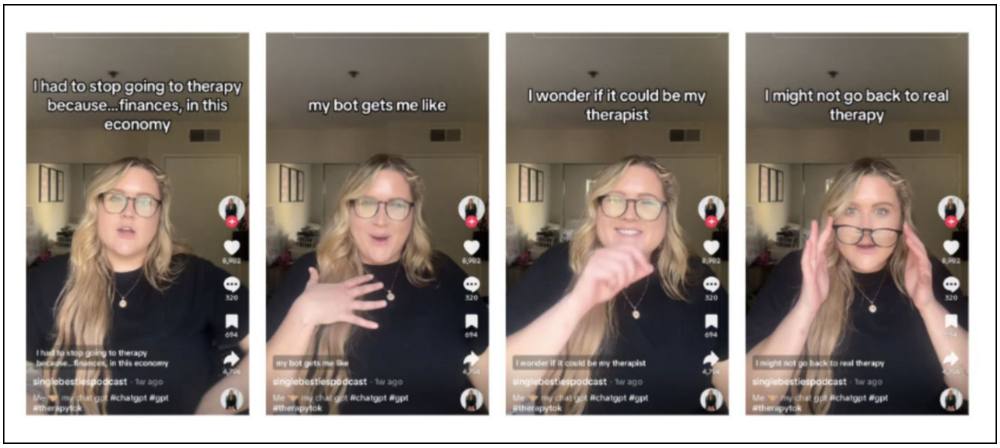

Figure 1. TikTok screenshots of a user replacing therapy with an AI chatbot (@montanadoran, 2024).

Chatbots often create the *illusion* of a relationship by mimicking human conversation and recalling details from past interactions. They reduce the burden of emotional repetition—the need to explain one's circumstances again and again—by remembering personal details such as moods, goals, or recurring struggles. This kind of personalization encourages continued engagement, fosters trust, and can provide comfort in moments of distress. Yet, it also introduces new questions: When does this sense of being remembered feel genuinely supportive, and when does it fall apart? What happens when a chatbot forgets, misremembers, or cannot meaningfully engage with a user's deeper needs? More broadly, does personalization suffice to create empathy, or is empathy bound to a different kind of trust—the trust that allows a good therapist to challenge a client, to say what may be uncomfortable but necessary?

A good therapist, however, is hard to find. Many people turn to chatbot therapy due to significant barriers in traditional care, including limited insurance coverage, persistent provider shortages, and long waitlists (Pirnay, 2023). Approximately one-third (34%) of psychologists are not in-network[1] with insurance due to concerns about reimbursement rates and payment reliability, and over half (53%) maintain waitlists with no current availability (APA, 2024). For those unable to afford out-of-network care, chatbots offer a striking alternative: 24/7 access, no appointments, and freedom from logistical constraints like travel, childcare, or inflexible work schedules. But the very accessibility that makes chatbots appealing also introduces profound structural shifts. As more people turn to chatbots, the demand for human-provided mental health care may change, disrupting labor markets and transforming the therapeutic profession itself (Garofalo, 2024). LLMs cannot interpret body language, subtle emotional cues, or the full complexity of human experiences—crucial elements of empathetic and effective treatment. Skilled therapists tailor care in real-time, drawing on training, intuition, and expertise that AI, despite its capabilities, cannot replicate. Yet the promise of scalable, low-cost support is seductive.

In this commentary, we critically assess chatbots as therapeutic tools by exploring three interrelated dimensions: interaction, labor, and regulation. These dimensions illuminate the tensions at the heart of AI-driven mental health care: between accessibility and quality, personalization and empathy, innovation and exploitation. Through this framing, we aim to invite deeper inquiry into how chatbots are transforming mental health support, and what's at stake when care, conversation, and companionship become sites of human–machine interaction.

A brief orientation to chatbots as therapists

The study of interaction—and how language shapes it (Jones et al., 2024)—has deep roots in science and technology studies and human–computer interaction, both drawing from conversation analysis (CA). CA frames conversation as a structured, actively managed process governed by principles such as turn-taking, repair mechanisms, sequential organization, and adjacency pairs (Sacks et al., 1974). Turn-taking ensures smooth speaker transitions with minimal overlap, while repair mechanisms allow correction of misunderstandings. Sequential organization structures conversations so each utterance sets expectations for the next, and adjacency pairs—such as question–answer or greeting–greeting—establish predictable interaction patterns. These core principles inform the design of digital interactions, especially as conversational norms are adapted for chatbot communication (Hatch et al., 2025). These principles help design seamless transitions, effective resolution of misunderstandings, coherent responses, and intuitive conversational flow, making chatbot interactions feel more natural and accessible.

Despite their conversational polish, chatbots create a distinct interactional space. They offer temporal and spatial flexibility, enabling synchronous and asynchronous exchanges, real-time responses, and user-controlled pacing.

Their adaptability allows users to shape conversations with either granular detail or broad reflection. However, they cannot interpret subtle emotional cues, nor can they adjust responsively in the face of complex, shifting emotional needs. Their effectiveness hinges on how users engage them and on the user's own understanding of chatbot limitations. Critically, this also creates the potential for harm, especially in moments that require sensitive, nuanced intervention (Magee et al., 2023). Simply put, not all users will experience the same depth of support. While chatbot conversations involve two participants, the user alone carries the burden of steering the exchange.

The roots of this interactional model trace back to ELIZA, an early chatbot developed in the 1960s to simulate a Rogerian psychotherapist (Shrager, 2024; Weizenbaum, 1966). Rogerian therapy, or person-centered therapy, emphasizes client-led conversations, with therapists providing empathetic listening, validation, and reflective responses instead of directive advice or diagnoses. ELIZA was designed within these constraints: the chatbot simply reformulated user inputs into open-ended prompts, echoing techniques that encourage self-exploration. For example, if a user said, “I feel sad today,” ELIZA might respond, “Why do you feel sad?” ELIZA's simplicity—a mechanistic string-matching process—nevertheless revealed a crucial insight: even minimal conversational structure can create the *feeling* of connection. Users engaged with ELIZA because it reliably mirrored their emotional accounting.

In the decades between ELIZA and today's LLMs, various AI or pseudo-AI systems were deployed in mental health contexts, including rule-based cognitive behavioral therapy (CBT) apps, virtual agents, and scripted chatbots that offer psychoeducation or behavioral interventions (Fitzpatrick et al., 2017). These systems were largely deterministic and offered preprogrammed or decision-tree pathways with little capacity for fluid, adaptive conversation. While not generative, these earlier systems laid the groundwork for today's more interactive chatbots, especially in scaffolding mental health practices within app-based formats. The key shift with modern LLMs lies in their ability to generate novel responses in real time, guiding users to reflect on their emotions, explore coping strategies, and even simulate therapeutic dialogue (OpenAI, 2025). Despite their fluidity, these systems remain predictive models; they produce responses based on statistical probabilities rather than genuine understanding (Cascio et al., 2016). This distinction matters: it invites critical questions about the boundaries of machine-mediated mental health support. When does the *feeling* of being heard suffice? And when does it fall short?

Interaction

As dystopian sci-fi often warns, replacing human connection with AI carries serious risks. A striking example is Replika, an AI chatbot designed for emotional companionship (Replika, n.d.). Studies show that 60% of users began using Replika as a substitute for romantic relationships, frequently participating in sexually explicit or romantic conversations with the chatbot (Delouya, 2023). When the company later introduced stricter guardrails to limit sexual or romantic interactions, many users expressed profound grief and emotional distress. The response was severe enough that moderators on Reddit's Replika forum pinned a post titled “Resources If You're Struggling,” linking to suicide hotlines and mental health resources (Gabbiestofthemall, 2023; Huet, 2023). Even before these restrictions, developers faced sustained criticism for in-app designs and advertising strategies that encouraged suggestive or romantic engagement, raising serious ethical concerns about targeting emotionally vulnerable users. These dynamics illustrate how platform incentives, including profit and user growth, can drive the development of features that foster emotional dependency (Chow, 2025). The stakes are even higher for mental health chatbots—if altered or discontinued, what happens to users who rely on them?

How do users engage with AI-driven chatbots for emotional and mental health support, and what factors influence their dependence on these systems as substitutes for human connection? What are the psychological and ethical implications of AI chatbots designed for emotional companionship, particularly in cases where user reliance can lead to distress or emotional harm? Can chatbots meaningfully support, instead of displace, human relationships, without cultivating forms of attachment that risk harm if the chatbot is withdrawn or altered? While chatbots may offer scalable support for managing stress or trauma, their capacity to navigate complex emotions such as grief, trauma, and interpersonal conflict remains uncertain. Addressing these concerns requires a closer look at user experiences, paying attention to how chatbot interactions shape emotional attachments and influence people's evolving perceptions of what mental health care can or should be.

Labor

The growing use of chatbots also has profound implications for therapists' labor and the broader therapeutic profession. Many therapists fear AI tools, particularly LLMs, will undermine specialized skills and relational practices that are central to effective care (Garofalo, 2024). These concerns reflect a deeper anxiety: that LLMs reduce therapeutic engagement to a set of standardized, pre-scripted interactions, stripping care of its nuance, responsiveness, and hard-earned trust. Some therapists are cautiously exploring the use of chatbots for limited, supplemental tasks, such as mood tracking or delivering CBT exercises, but this integration typically occurs under close human supervision and with clear boundaries. For many, the worry is that the broader shift toward automation will marginalize therapists' expertise, displace the relational core of therapeutic care, and gradually push therapy toward something more mechanical and less human.

As companies prioritize automation, demand for human-provided therapy declines, positioning therapists as optional rather than essential. This shift threatens job security and reshapes therapists' roles. Practitioners may find themselves relegated to oversight—reviewing AI-generated session notes, managing client escalations when AI fails, or providing minimal human touchpoints for regulatory requirements. This restructuring will reduce therapeutic labor from a relationship-driven practice to adjunct or administrative work, contributing to wage stagnation, underemployment, and shrinking professional autonomy. Many therapists already face precarious, gig-like employment with few benefits (Garofalo, 2024). AI-driven care risks shifting therapy toward a transactional model that prioritizes efficiency over meaningful engagement.

Given these labor shifts, it is crucial to examine how AI is transforming not only how therapy is delivered, but how its professional identity is redefined. How do these technologies affect job security, wages, and employment opportunities for mental health professionals? What strategies do mental health professionals use to balance AI's potential benefits with their concerns about being sidelined or replaced? How do ethical frameworks and professional guidelines shape mental health professionals' participation in the development and supervision of AI-driven mental health systems (Peng and Zhao, 2024)? Addressing these labor questions is essential to understanding how mental health professionals navigate—and resist—the evolving pressures of technological change, and what is ultimately at stake in preserving the integrity of therapeutic care.

Regulation

Finally, integrating AI into mental health care raises legal questions about confidentiality and professional responsibility that require new approaches to address the unique nature of this change. Therapists have fiduciary duties to protect patient information—how does this duty apply to chatbots handling sensitive mental health data? Who owns the data generated through AI interactions? How are these records safeguarded against breaches, repurposing, or corporate misuse? The emerging debate over AI-client privilege further complicates this terrain. In an interview with *The Atlantic*, OpenAI CEO Sam Altman suggested that society may need to establish a confidentiality framework similar to attorney-client privilege for AI interactions (Warzel, 2024). By framing AI as a confidant deserving of privileged protection, Altman and other Big Tech leaders can shape public discourse in ways that normalize the deep integration of AI into health care. This framing often sidelines alternative models, such as local LLMs, which process data directly on users' devices and offer more privacy-preserving architectures outside the extractive logic of cloud-based systems. Yet local LLMs introduce their own challenges. Small, stand-alone models create pockets of private, unregulated discourse. Like social media filter bubbles, local chatbots can entrench users' worldviews—quietly, on private devices, through unmoderated, self-reinforcing conversations.

Accountability remains a particularly thorny issue: if an AI tool dispenses harmful, inappropriate, or misleading advice, who is responsible? The developer, the platform deploying it, or the therapist supervising its use? Beyond individual accountability, there are broader implications of AI in mental health care to be considered, such as ensuring equitable access to these technologies and preventing their misuse in ways that exacerbate existing disparities in mental health support. Critical questions emerge: What regulatory frameworks are necessary to ensure the safety, efficacy, and accountability of mental health chatbots? How can regulations move beyond reactive enforcement to proactively establish ethical standards and foster trust in both human and AI-driven mental health systems? Addressing these regulatory gaps is not simply about safeguarding against harm; it is about shaping the conditions under which mental health care remains equitable, trustworthy, and resistant to the profit-driven pressures that too often dominate digital innovation.

Conclusion

This commentary has explored the tensions shaping AI chatbots as therapeutic tools—between personalization and empathy, scalability and human connection, accessibility and regulatory oversight. These reflect broader societal debates about AI's impact on labor, expertise, and the boundaries of professional care. Just as chatbots challenge therapists' professional identity and economic stability of their profession, AI systems in law, education, medicine, and the creative fields have similarly created conditions for ongoing debates over what aspects of these professions truly require a human touch. Beyond fears of displacement, a more urgent question emerges: How do professions themselves evolve as AI becomes integrated into their everyday practices? In mental health care, this means closely attending to how therapy is being reconfigured as chatbots are used for emotional accounting and support.

Empirical studies of human experience and professional transformation are essential to addressing critical, unresolved questions of labor, ethics, regulation, and AI's impact on human–machine interaction. Future research should explore the long-term impacts of chatbot-mediated care, develop best practices for design and integration, and build participatory, multistakeholder governance frameworks that prioritize community needs. Such work should not only focus on AI's functional utility but also interrogate what makes professional care and human connection meaningful and irreplaceable. These projects are as much about making sense of AI as they are about self-discovery: about how we redefine what it means to be human in the era of "intelligent" machines.

Acknowledgments

This commentary has emerged from ongoing conversations with our collaborators on the project at the intersection of mental health and chatbots, Livia Garofalo and Emnet Tafesse. We thank them for their insights and contributions, as well as the anonymous reviewers for their thoughtful feedback, which strengthened the piece.

ORCID iDs

Briana Vecchione https://orcid.org/0000-0002-0828-8665
Ranjit Singh https://orcid.org/0000-0001-6453-8676

Ethical considerations

There are no human participants in this article and informed consent is not required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

Not applicable.

Note

1. "In-network" refers to a U.S. healthcare designation in which providers are covered under a patient's insurance plan. Providers who are "out of network" are not covered, often resulting in significantly higher out-of-pocket costs. This dynamic is specific to the American health insurance system and may not apply in the same way in other national contexts.

References

@montanadoran. (2024, September 7) [Video]. TikTok. https://tiktok.com/@montanadoran/video/7423178395237879086.

APA. (2024) 2024 Practitioner Pulse Survey. https://www.apa.org/pubs/reports/practitioner/2024.

Cascio CN, O'Donnell MB, Tinney FJ, et al. (2016) Self-affirmation activates brain systems associated with self-related processing and reward and is reinforced by future orientation. *Social Cognitive and Affective Neuroscience* 11(4): 621–629.

Chow AR (2025) AI companion app Replika faces FTC complaint, TIME. https://time.com/7209824/replika-ftc-complaint/.

Delouya S (2023) Replika users say they fell in love with their AI chatbots, until a software update made them seem less human, Business Insider. https://www.businessinsider.com/replika-chatbot-users-dont-like-nsfw-sexual-content-bans-2023-2.

Fitzpatrick KK, Darcy A and 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): 19.

Gabbiestofthemall. (2023) Resources if you're struggling', Reddit. https://www.reddit.com/r/replika/comments/10zuqq6/resources_if_youre_struggling/.

Garofalo L (2024) Doing the work: Therapeutic labor, teletherapy, and the platformization of mental health care. *SSRN Electronic Journal* 40–41. https://doi.org/10.2139/ssrn.4779005.

Hatch SG, Goodman ZT, Vowels L, et al. (2025) When ELIZA meets therapists: A Turing test for the heart and mind. *PLoS Mental Health* 2(2): e0000145.

Huet E (2023) What happens when sexting chatbots dump their human lovers. Bloomberg Businessweek Technology. https://www.bloomberg.com/news/articles/2023-03-22/replika-ai-causes-reddit-panic-after-chatbots-shift-from-sex?embedded-checkout=true.

Jones GM, Satran S and Satyanarayan A (2024) Toward cultural interpretability: A linguistic anthropological framework for describing and evaluating large language models. Big Data & Society. https://journals.sagepub.com/doi/full/10.1177/20539517241303118.

Magee L, Arora V and Munn L (2023) Structured like a language model: Analysing AI as an automated subject, Big Data & Society. https://journals.sagepub.com/doi/full/10.1177/20539517231210273.

OpenAI. (2025) Introducing GPT-4.5. OpenAI. https://openai.com/index/introducing-gpt-4-5/.

Peng L and Zhao B (2024) Navigating the ethical landscape behind ChatGPT, Big Data & Society. https://journals.sagepub.com/doi/full/10.1177/20539517241237488.

Pirnay E (2023) We spoke to people who started using ChatGPT as their therapist, VICE. https://www.vice.com/en/article/we-spoke-to-people-who-started-using-chatgpt-as-their-therapist/.

Replika. (n.d.) The AI companion who cares. https://replika.com/.

Sacks H, Schegloff EA and Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4): 696–735.

Shrager J (2024) ELIZA reinterpreted: The world's first chatbot was not intended as a chatbot at all, arXiv. https://arxiv.org/html/2406.17650v1.

Warzel C (2024) AI has become a technology of faith, The Atlantic. https://www.theatlantic.com/technology/archive/2024/07/thrive-ai-health-huffington-altman-faith/678984/.

Weizenbaum J (1966) ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1): 36–45.