MATHEWS

Review Article

Mathews Journal of Psychiatry & Mental Health

Simulated Empathy and Therapeutic Mechanisms in AI-Based Mental Health Chatbots: A Critical Narrative Review

Tamana A Bhat*, Rakesh Kumar Mishra

Amity Institute of Psychology and Allied Sciences, Amity University Noida, India

ABSTRACT

Psychological distress, including anxiety, depression, chronic stress, and loneliness remains a major global public health concern. AI-based mental health chatbots are increasingly proposed as scalable, low-cost interventions. This critical narrative review examines whether the therapeutic mechanisms of such chatbots, particularly simulated empathy and structured cognitive- behavioral techniques, can meaningfully replicate or appropriately adapt the active ingredients of psychotherapy. Drawing on cognitive-behavioral and humanistic frameworks (e.g., Rogers' core conditions), we argue that while AI systems may activate structured coping mechanisms and enhance short-term engagement, the absence of affective reciprocity and relational depth places inherent limits on the scope of change they can facilitate. Clinical and ethical implications are discussed, with particular attention to stepped-care models, accountability, and the Indian public health context.

Keywords: Artificial Intelligence, Mental Health Chatbots, Simulated Empathy, Therapeutic Alliance, Cognitive Behavioral Therapy (CBT), Digital Mental Health, Emotional Support, Psychotherapy Mechanisms, Stepped-Care Model, Ethical Implications, Accessibility, India Context

INTRODUCTION

Psychological distress, including anxiety, depression, chronic stress, and loneliness, represents a major global public health concern. Beyond emotional suffering, these conditions significantly impair interpersonal functioning, occupational performance, and overall quality of life. Despite increased awareness, mental health care systems remain unable to meet the growing demand for services. Structural barriers such as financial cost, shortages of trained clinicians, geographical inequities, stigma, and long waiting lists continue to limit access to evidence-based treatment [1]. These limitations have intensified interest in scalable, low-cost interventions that can extend psychological support beyond traditional face-to-face therapy.

Vol No: 11, Issue: 01

Received Date: April 15, 2026

Published Date: April 29, 2026

***Corresponding Author**

Tamana A Bhat,

Amity Institute of Psychology and Allied Sciences, Amity University Noida, 125, Noida, Uttar Pradesh 201301, India,

Phone: +91-7889675075,

ORCID: 0009-0001-9917-6545;

Emails: tamana.ajaz@s.amity.edu; tamanaajaz7@gmail.com

Citation: Bhat TA, et al. (2026). Simulated Empathy and Therapeutic Mechanisms in AI-Based Mental Health Chatbots: A Critical Narrative Review. Mathews J Psychiatry Ment Health. 11(1):62.

One such development is the rise of AI-based mental health chatbots, which use natural language processing and machine learning to simulate therapeutic dialogue. Many of these systems are explicitly grounded in structured, skills-based interventions, particularly cognitive behavioral therapy (CBT), behavioral activation, and mindfulness-based approaches [2,3]. Applications such as Woebot and Wysa aim to deliver psychoeducation, cognitive restructuring exercises, mood monitoring, and coping strategies through conversational interfaces. From a stepped-care perspective, these tools are positioned as low-intensity interventions that may increase access, reduce stigma-related avoidance, and provide early support before symptoms escalate [4].

However, their clinical relevance depends not only on accessibility but also on whether they meaningfully replicate or appropriately adapt the active ingredients of psychotherapy.

A central question concerns therapeutic mechanisms. Within CBT, symptom change is theorized to occur through structured cognitive and behavioral processes: identifying maladaptive thoughts, testing cognitive distortions, increasing adaptive behaviors, and reinforcing self- efficacy. Chatbots are particularly well-suited to delivering these structured components because CBT is manualized, directive, and skills-focused. Automated systems can guide users through thought records, behavioral activation tasks, and mood tracking with consistency and scalability. Randomized controlled trials suggest that such interventions produce small to moderate reductions in depressive and anxiety symptoms [2,4]. Analytically, this raises the possibility that for mild to moderate distress, structured cognitive-behavioral techniques may account for a substantial proportion of therapeutic benefit, potentially reducing reliance on relational factors at least in the short term.

However, psychotherapy research has long emphasized that therapeutic change is not driven solely by technique. Within humanistic theory, Carl Rogers proposed that psychological growth occurs when clients experience specific "core conditions" within the therapeutic relationship: empathy, unconditional positive regard, and congruence [5]. Rogers conceptualized empathy as an active, moment-to-moment effort to understand the client's internal frame of reference "as if" one was the client, without losing the therapist's own grounded position. This form of empathy is not merely cognitive accuracy but involves affective attunement and relational presence. Later process research further supported the importance of perceived empathic understanding as a predictor of therapeutic outcome [6].

AI-based chatbots, in contrast, generate empathic responses through algorithmic detection of emotional cues and scripted validation. Although this may approximate *cognitive* empathy (accurate recognition and reflection of expressed affect), it does not involve affective resonance or reciprocal interpersonal experience. Thus, chatbots reproduce the linguistic form of empathy without its intersubjective depth. The theoretical tension lies in whether perceived validation alone is sufficient to facilitate engagement and symptom reduction, or whether the absence of authentic relational presence constrains the scope of change. This distinction becomes particularly salient in relation to the therapeutic alliance, which encompasses agreement on goals, collaboration on tasks, and the relational bond. Chatbots may effectively support goal and task alignment through structured guidance, yet the bond component remains conceptually limited. While users frequently report feeling supported and understood [7,8], this experience reflects algorithmic responsiveness rather than intersubjective exchange.

USE OF GENERAL-PURPOSE CONVERSATIONAL AI FOR EMOTIONAL SUPPORT

In addition to tools specifically designed around therapeutic frameworks, many individuals are increasingly using general-purpose conversational AI systems such as ChatGPT and Google Bard for emotional support. Although these systems were not developed as clinical interventions, emerging research suggests that users frequently engage them for emotional expression, stress management, reassurance seeking, and cognitive reframing [8,9]. Survey-based findings indicate particularly high engagement among adolescents and young adults, populations already accustomed to digital modes of self-disclosure and help-seeking [10,11]. The appeal of these systems can be understood through models of digital self-disclosure and reduced stigma in online environments [12].

From a mechanistic perspective, many user-AI interactions appear to overlap with cognitive- behavioral processes. Users commonly request help in identifying distorted thinking patterns, generating alternative interpretations, or developing coping strategies mechanisms central to CBT's theory of change [13]. However, in the context of general-purpose AI, these processes occur without formal assessment, case conceptualization, or systematic monitoring.

Consequently, while AI may facilitate isolated cognitive shifts, it does not deliver therapy within a coherent treatment model. When individuals with moderate to severe psychological distress substitute AI interaction for professional care, there is potential for delayed treatment, insufficient risk detection, and overreliance on non-clinical guidance [8].

ACCESSIBILITY AND PUBLIC MENTAL HEALTH IMPLICATIONS OF CONVERSATIONAL AI

AI-based chatbots can offer meaningful, short-term emotional support for individuals experiencing mild to moderate psychological distress. Chatbot interventions grounded in evidence-based approaches such as CBT and mindfulness are associated with modest reductions in symptoms of anxiety, depression, stress, and loneliness [2,4]. Although these effects are generally smaller than those achieved through therapist-led psychotherapy, they remain clinically meaningful within a public mental health framework that prioritizes reach, scalability, and early intervention.

The relevance of these findings becomes particularly pronounced in low- and middle-income countries such as India, where the treatment gap for mental disorders remains substantial.

National data from the National Mental Health Survey of India indicate that treatment gaps for certain conditions exceed 70% [14]. Contributing factors include shortages of trained mental health professionals, urban-rural disparities, financial barriers, stigma, and limited mental health literacy. In this context, conversational AI may hold public health relevance. India has experienced rapid growth in smartphone penetration and internet accessibility, including in semi-urban and rural regions. Low-cost, mobile-based AI tools could therefore function as scalable, low-intensity supports within a stepped-care model, offering early coping assistance for individuals with mild distress who might otherwise receive no intervention. However, digital literacy varies significantly, and linguistic diversity presents challenges for culturally and contextually sensitive AI responses. Psychological distress in India is often embedded within complex sociocultural stressors family dynamics, academic pressure, unemployment, gender- based constraints, and collectivistic value systems that require nuanced, culturally informed understanding. AI systems lack the cultural attunement and relational negotiation often required in Indian therapeutic settings.

CLINICAL AND ETHICAL IMPLICATIONS

The integration of conversational AI into mental health support systems carries significant clinical and ethical implications that extend beyond questions of efficacy to encompass fundamental considerations of safety, accountability, and appropriate scope of practice.

Clinically, the evidence supports positioning AI chatbots as low-intensity, adjunctive tools within stepped-care models rather than as replacements for psychotherapy [4,8]. Their demonstrated strengths lie in delivering structured, skills-based interventions that align with the technical components of CBT. For individuals experiencing mild to moderate distress who might otherwise receive no intervention, these tools can provide accessible, immediate coping support. However, clinical appropriateness diminishes as psychological complexity increases.

Psychotherapy involves more than technique delivery; treatment outcomes are strongly associated with relational factors that AI cannot replicate, including the capacity to navigate interpersonal ruptures and provide genuine affective attunement [15]. A chatbot cannot conduct a suicide risk assessment, formulate a safety plan responsive to dynamic risk factors, or make discretionary clinical judgments that differentiate competent practice from protocol-driven response.

Ethically, conversational AI operates outside the professional accountability frameworks that govern clinical practice. Licensed clinicians are bound by enforceable codes of conduct, continuing education requirements, and liability structures that provide recourse for harm [16]. AI systems, by contrast, lack moral agency, professional liability, and discretionary judgment. This creates significant ambiguity

in situations involving risk disclosure: if a user expresses suicidal ideation to a chatbot, what obligations exist? Who is responsible the developer, the platform, or the algorithm itself? Current regulatory frameworks provide inconsistent answers. Data privacy represents a further critical concern. Mental health disclosures involve inherently sensitive information, and users may share intimate details with AI systems under the assumption of confidentiality [8]. Yet data governance mechanisms vary widely across platforms, and commercial interests may conflict with clinical confidentiality principles. Transparent consent procedures, clear data usage policies, and robust security safeguards are therefore essential prerequisites for ethical deployment. In contexts such as India, where substantial treatment gaps persist [14], AI tools present both opportunity and risk, requiring clear scope definition, embedded referral pathways, cultural adaptation, and regulatory alignment.

CLINICAL POSITIONING

Clinically, the evidence supports positioning AI chatbots as low-intensity, adjunctive tools within stepped-care models rather than as replacements for psychotherapy. Their demonstrated strengths lie in delivering structured, skills-based interventions, cognitive restructuring, behavioral activation, and psychoeducation that align with the technical components of cognitive behavioral therapy [13]. For individuals experiencing mild to moderate distress who might otherwise receive no intervention, these tools can provide accessible, immediate coping support. However, clinical appropriateness diminishes as psychological complexity increases.

Psychotherapy involves more than technique delivery; treatment outcomes are strongly associated with relational factors that AI cannot replicate. The therapeutic alliance, for instance, encompasses not only agreement on goals and tasks but also a relational bond characterized by mutual trust, affective connection, and the capacity to navigate interpersonal ruptures [15]. While chatbots can support goal and task alignment through structured guidance, the bond component remains conceptually and practically inaccessible to algorithmic systems.

Users may feel a connection, but this reflects what Turkle S [17] terms “the illusion of companionship without the demands of friendship” a simulated relationship that may satisfy momentary needs but does not constitute genuine interpersonal engagement. Similarly, although users frequently report feeling understood by AI systems, such empathy is computationally generated rather than affectively experienced. Drawing again on the reports versus affective empathy distinction: AI can recognize emotional keywords and generate validating responses, but it cannot engage in the moment-to-moment attunement, implicit communication, or reciprocal emotional resonance that characterizes human empathic connection [5,6]. Consequently, its use in moderate to severe psychopathology particularly in cases involving suicidality, trauma, or complex comorbidity raises concerns regarding clinical adequacy and safety. A chatbot cannot conduct a suicide risk assessment, formulate a safety plan responsive to dynamic risk factors, or make the discretionary clinical judgments that differentiate competent practice from protocol-driven response.

ETHICAL ACCOUNTABILITY

Ethically, conversational AI operates outside the professional accountability frameworks that govern clinical practice. Licensed clinicians are bound by enforceable codes of conduct, continuing education requirements, and liability structures that provide recourse for harm [16]. AI systems, by contrast, lack moral agency, professional liability, and discretionary judgment. This creates significant ambiguity in situations involving risk disclosure: if a user expresses suicidal ideation to a chatbot, what obligations exist? Who is responsible the developer, the platform, the algorithm itself? Current regulatory frameworks provide inconsistent answers. Data privacy represents a further critical concern.

Mental health disclosures involve inherently sensitive information, and users may share intimate details with AI systems under the assumption of confidentiality [8]. Yet data governance mechanisms vary widely across platforms, and commercial interests may conflict with clinical confidentiality principles. Transparent consent procedures, clear data usage policies, and robust security safeguards are therefore essential prerequisites for ethical deployment.

CONTEXTUAL CONSIDERATIONS

In contexts such as India, where substantial treatment gaps persist [14], AI tools present both opportunity and risk. The opportunity lies in expanding access to early-

stage support for populations underserved by traditional services. The risk involves over-reliance on tools that cannot provide comprehensive care, potentially delaying necessary professional intervention when symptoms escalate. Ethical implementation therefore requires:

1. ***Clear scope definition:*** Public communication about what AI can and cannot do
2. ***Embedded referral pathways:*** Seamless connections to human clinicians when risk indicators emerge
3. ***Cultural adaptation:*** Responsiveness to linguistic diversity and sociocultural contexts
4. ***Regulatory alignment:*** Compliance with national mental health and data protection frameworks

DISCUSSION

The current evidence indicates that conversational AI can produce small to moderate short-term reductions in symptoms of anxiety, depression, and stress, particularly when interventions are grounded in structured cognitive-behavioral principles [2,4]. These findings, however, require careful contextualization. Most studies examine brief interventions targeting mild to moderate symptom presentations, with limited follow-up periods. The data therefore support AI chatbots as low-intensity, early-stage supports rather than comprehensive therapeutic interventions.

Mechanistically, the observed benefits appear primarily driven by what psychotherapy research identifies as technical rather than relational factors. The structured components of CBT cognitive restructuring exercises, behavioral activation prompts, psychoeducation, and self-monitoring can be effectively operationalized within algorithmic frameworks [16]. Chatbots excel at delivering these manualized elements with consistency and scalability. However, where psychological complexity intensifies, the limitations of an algorithmic model become increasingly pronounced. The therapeutic alliance requires not only goal and task alignment but also a relational bond characterized by mutual trust, affective connection, and the capacity to navigate interpersonal ruptures [9]. While chatbots can support goal and task alignment through structured guidance, the bond component remains conceptually and practically inaccessible to algorithmic systems.

Users may feel a connection, but this reflects what Turkle [17] terms "the illusion of companionship without the demands of friendship" a simulated relationship that may satisfy momentary needs but does not constitute genuine interpersonal engagement.

The appropriate integration of conversational AI, therefore, is neither dismissal nor uncritical adoption, but calibrated placement within stepped-care frameworks. As adjunctive, low-intensity supports, they extend reach. As replacements for psychotherapy, they risk fundamentally misrepresenting both the nature of psychological distress and the mechanisms of meaningful change. Future research must move beyond short-term symptom metrics to examine durability, functional outcomes, and unintended consequences such as delayed help-seeking [8]. In sum, conversational AI can activate coping processes through structured guidance, but it cannot replicate the relational depth that constitutes the heart of psychotherapeutic change. It is the presence of another mind not the output of an algorithm that ultimately transforms suffering into growth.

CONCLUSION

A very high percentage of the population experiences mild to moderate psychological dysfunction, and CBT-based interventions delivered by artificial intelligence could partially help both patients and therapists. Let us not forget that therapies and work often fail because they become tainted by human emotions such as jealousy, envy, and resentment. Artificial intelligence removes those contaminating factors and is immensely useful as a low-intensity, adjunctive tool. However, its limitations must be clearly communicated, and ethical safeguards including referral pathways, cultural adaptation, and regulatory alignment must be implemented to prevent over-reliance and ensure patient safety.

ACKNOWLEDGEMENTS

None.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Kazdin AE, Blase SL. (2011). Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. Perspect Psychol Sci. 6(1):21-37.
2. Fitzpatrick KK, 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 Ment Health. 4(2):e19.
3. 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 Uhealth. 6(11):e12106.
4. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. (2019). An overview of the features of chatbots in mental health: A scoping review. Int J Med Inform. 132:103978.
5. Rogers CR. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology. 21(2):95-103.
6. Elliott R, Bohart AC, Watson JC, Greenberg LS. (2011). Empathy. Psychotherapy (Chic). 48(1):43-49.
7. Ho A, Hancock J, Miner AS. (2018). Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations with a Chatbot. J Commun. 68(4):712-733.
8. Miner AS, Milstein A, Hancock JT. (2017). Talking to Machines About Personal Mental Health Problems. JAMA. 318(13):1217-1218.
9. Bickmore TW, Schulman D, Sidner C. (2013). Automated interventions for multiple health behaviors using conversational agents. Patient Educ Couns. 92(2):142-148.
10. Pretorius C, Chambers D, Coyle D. (2019). Young People's Online Help-Seeking and Mental Health Difficulties: Systematic Narrative Review. J Med Internet Res. 21(11):e13873.
11. Rideout V, Fox S. (2018). Digital health practices, social media use, and mental well-being among teens and young adults. Hopelab, USA. pp.1-96.
12. Andalibi N, Ozturk P, Forte A. (2018). Sensitive self-disclosures, responses, and social support on Instagram: The case of #depression. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW). pp. 1-22.
13. Beck JS. (2011). Cognitive behavior therapy: Basics and beyond (2nd ed.). Guilford Press, USA.
14. Gururaj G, Varghese M, Benegal V, Rao GN, Pathak K, Singh LK, et al. (2016). National Mental Health Survey of India, 2015-16: Summary. National Institute of Mental Health and Neurosciences. NIMHANS Publication No. 129.
15. Wampold BE, Imel ZE. (2015). The great psychotherapy debate: The evidence for what makes psychotherapy work (2nd ed.). Routledge, United Kingdom.
16. American Psychological Association. (2017). Ethical principles of psychologists and code of conduct. American Psychological Association, USA.
17. Turkle S. (2011). Alone together: Why we expect more from technology and less from each other. Basic Books.