International Journal of Humanities and Social Science Research
www.socialsciencejournal.in
ISSN: 2455-2070
Received: 21-04-2026, Accepted: 22-05-2026, Published: 06-06-2026
Volume 12, Issue 2, 2026, Page No. 284-287

Artificial intelligence in mental health care: Opportunities, ethical challenges, and the path forward

Asim Ahmed Laskar
Department of Sociology, Silchar Jadunath Sarkar School of Social Sciences, Assam University, Silchar, Assam, India
DOI: https://doi.org/10.66856/ijhssr.2026.12.2.12201

Abstract
The global mental health crisis has intensified demand for scalable, accessible, and evidence-based psychological interventions. Artificial intelligence (AI) technologies — including machine learning algorithms, natural language processing (NLP), and large language models (LLMs) — have emerged as potentially transformative tools in the assessment, diagnosis, and treatment of mental health conditions. This paper provides a comprehensive review of the current landscape of AI applications in mental health care, examining empirical evidence for their efficacy, systemic barriers to adoption, and the ethical tensions that accompany their deployment. Drawing on peer-reviewed literature, clinical case studies, and emerging regulatory frameworks, the analysis reveals a complex picture: AI holds genuine promise for broadening access to care and enhancing diagnostic precision, yet also risks exacerbating existing health disparities, eroding therapeutic alliance, and generating opaque algorithmic decisions with life-altering consequences. The paper concludes by proposing a human-centered integration framework grounded in clinical oversight, algorithmic transparency, and equity-centered design principles.

Keywords: Artificial intelligence, mental health, machine learning, natural language processing, digital therapeutics, algorithmic bias, therapeutic alliance, health equity

Introduction

Mental illness represents one of the most pressing public health challenges of the twenty-first century. The World Health Organization estimates that approximately one billion individuals worldwide live with a mental or neurological disorder, while fewer than half receive adequate treatment (World Health Organization [WHO], 2022) [23]. Treatment gaps are most acute in low- and middle-income countries, where the ratio of mental health workers to population can be as low as one per 100,000 individuals (Patel *et al*., 2018) [17]. Even in high-income nations, long waitlists, geographic barriers, and stigma conspire to prevent millions from accessing timely care.

Against this backdrop, the rapid proliferation of digital technologies and AI has generated substantial interest among clinicians, researchers, and policymakers. AI systems — broadly defined as computational systems capable of performing tasks that would ordinarily require human intelligence — have demonstrated notable performance on specific mental health-related tasks, from detecting depressive symptomatology in social media text (Coppersmith *et al*., 2015) [4] to predicting suicide risk from electronic health records (Walsh *et al*., 2017) [22]. Conversational AI agents and smartphone-based digital therapeutics have reached millions of users, often in contexts where professional care is unavailable (Torous *et al*., 2021) [21].

Yet enthusiasm must be tempered with critical scrutiny. The integration of AI into mental health care raises profound questions about clinical validity, algorithmic fairness, data privacy, and the irreplaceable role of the therapeutic relationship. This paper seeks to map the contemporary landscape of AI in mental health, evaluate the evidence base, and advance a framework for responsible integration that centers both clinical efficacy and ethical accountability.

Background and Conceptual Framework

1. Defining AI in the Mental Health Context

The term "artificial intelligence" encompasses a broad family of computational techniques. Within mental health applications, three categories predominate. First, supervised machine learning (ML) algorithms — including support vector machines, random forests, and deep neural networks — are trained on labeled datasets to classify or predict clinical outcomes, such as diagnostic category or treatment response (Shatte *et al*., 2019) [19]. Second, natural language processing (NLP) enables systems to extract clinically meaningful information from unstructured text, including therapy transcripts, clinical notes, and social media data (Gkotsis *et al*., 2017) [7]. Third, large language models (LLMs), trained on vast corpora of text, can engage users in extended, contextually coherent conversation and have been deployed in therapeutic chatbot applications (Abd-Alrazaq *et al*., 2020) [1].

2. The Global Mental Health Treatment Gap

Understanding AI's potential requires situating it within the structural realities of global mental health service delivery. The treatment gap — defined as the proportion of individuals with a diagnosable condition who receive no treatment — is estimated at 76-85% for low-income countries and 35-50% for high-income countries (Kohn *et al*., 2004) [10]. Contributing factors include workforce shortages, geographic inaccessibility, financial costs, cultural stigma, and inadequate health system investment. Proponents of AI-based solutions argue that digital tools can bypass many of these structural barriers by providing low-cost, always-available, and destigmatized interventions at population scale.

Current Applications of AI in Mental Health

1. Screening and Diagnostic Support

One of the most actively researched applications of AI in mental health is the automation or augmentation of clinical

screening and diagnosis. Studies have demonstrated that ML algorithms trained on neuroimaging data can distinguish between individuals with and without major depressive disorder with accuracy exceeding conventional clinical assessment in controlled settings (Orrù *et al*., 2012) [16]. Similarly, NLP models applied to acoustic features of speech — including prosody, speaking rate, and vocal affect — have shown promise as biomarkers for conditions including depression, schizophrenia, and post-traumatic stress disorder (Cummins *et al*., 2015) [5].

Passive sensing paradigms, which use data from smartphone accelerometers, GPS trajectories, and usage patterns, offer a potentially unobtrusive method of longitudinal symptom monitoring. Saeb *et al*. (2015) [18] demonstrated that smartphone GPS and accelerometer data could predict PHQ-9 depression scores with moderate accuracy, raising the possibility of continuous, real-world monitoring that bypasses the limitations of episodic clinical encounters. However, the external validity of many such studies remains limited by small sample sizes and non-representative participant pools.

2. Conversational Agents and Digital Therapeutics

Perhaps the most visible and controversial AI application in mental health is the deployment of conversational agents — chatbots — designed to deliver therapeutic support. Woebot, one of the most widely studied examples, is a smartphone-based chatbot that delivers elements of cognitive-behavioral therapy (CBT) through natural language dialogue. A randomized controlled trial by Fitzpatrick *et al*. (2017) [6] found that Woebot users demonstrated significantly greater reductions in depression and anxiety symptoms over two weeks compared to a control condition, with high engagement rates among college-aged participants.

These findings are encouraging but require careful interpretation. The study population was predominantly young, educated, and already self-selected for interest in digital health tools. The intervention period was brief, and longer-term outcomes have not been established. Critics note that chatbots may be most appropriate as adjuncts to professional care — extending between-session support or providing psychoeducation — rather than as standalone treatments, particularly for individuals with moderate-to-severe psychopathology (Inkster *et al*., 2018) [9].

3. Predictive Analytics and Risk Stratification

Predictive analytics applications aim to identify individuals at elevated risk for adverse outcomes including self-harm, suicide attempts, and psychiatric hospitalization. Using electronic health record (EHR) data including medication prescriptions, diagnosis codes, and clinical notes, ML models have demonstrated sensitivity and specificity that compares favorably to clinician assessment in certain settings (Simon *et al*., 2018) [20]. The U.S. Department of Veterans Affairs has implemented one such system at scale, flagging veterans at elevated suicide risk for intensified outreach (McCarthy *et al*., 2015) [12].

The deployment of predictive risk tools raises acute ethical concerns, however. False positives may subject individuals to stigmatizing interventions or unwanted clinical attention, while false negatives may foster complacency in clinicians who over-rely on algorithmic guidance. The "black box" nature of many high-performing ML models renders their decision logic opaque, making clinical auditing and accountability difficult (Obermeyer & Emanuel, 2016) [14].

Ethical Challenges and Critical Considerations

1. Algorithmic Bias and Health Equity

A fundamental concern with AI systems trained on historical health data is the risk of perpetuating or amplifying existing disparities. Training datasets systematically underrepresent racial and ethnic minorities, individuals with lower socioeconomic status, and those in rural regions (Chen *et al*., 2021) [3]. When AI systems are subsequently deployed in these populations, their reduced accuracy may translate directly into differential diagnostic quality — effectively providing lower-quality care to those who already face the greatest structural barriers.

Obermeyer *et al*. (2019) [15] demonstrated this problem compellingly in a general health context: a widely used commercial algorithm systematically underestimated illness severity in Black patients relative to white patients, because it used health care costs as a proxy for health need — a variable that is itself confounded by historical discrimination. Analogous biases are likely present in mental health AI systems, and their detection requires rigorous disaggregated performance evaluation across demographic subgroups — a standard that current regulatory frameworks do not consistently mandate.

2. Privacy, Data Governance, and Informed Consent

Mental health data are among the most sensitive categories of personal information, carrying significant risks of harm if disclosed — including employment discrimination, stigmatization, and interpersonal consequences. AI mental health tools typically require continuous data collection from smartphones, wearables, and digital platforms, generating longitudinal behavioral profiles of considerable granularity. Many commercially deployed applications have been found to share user data with third parties, including advertisers, in ways that users may not fully comprehend when providing consent (Grundy *et al*., 2019) [8].

Meaningful informed consent in the context of AI systems presents unique challenges. Standard consent processes cannot fully communicate the scope of potential data uses, the nature of algorithmic inferences, or the risks of re-identification in nominally anonymized datasets. A more robust consent framework would require ongoing, revocable consent; clear disclosure of commercial data-sharing arrangements; and accessible explanations of how AI systems use personal data to generate outputs.

3. Therapeutic Alliance and the Human Element

The therapeutic relationship — characterized by empathic attunement, trust, rupture and repair, and the mutual recognition of shared humanity — is considered by many clinicians and researchers to be the central mechanism of psychotherapeutic change (Norcross & Wampold, 2011) [13]. The integration of AI into clinical encounters raises fundamental questions about whether, and to what degree, this relationship can be instantiated, simulated, or replaced by computational systems.

Users of AI-based therapeutic tools frequently report positive experiences, including feelings of being heard and supported. However, it remains unclear whether these perceived benefits reflect genuine therapeutic mechanisms or the projection of relational qualities onto a system designed to simulate them (Luxton, 2014) [11]. Clinicians express concern that AI-mediated interactions may inadvertently reinforce avoidance of the genuine interpersonal vulnerability that many therapeutic modalities aim to cultivate.

Emerging Regulatory and Professional Frameworks

The regulatory landscape governing AI in mental health care remains nascent and fragmented. In the United States, the Food and Drug Administration (FDA) has developed a framework for Software as a Medical Device (SaMD), distinguishing between tools that provide information for clinical decision support (lower regulatory burden) and those that drive or inform diagnosis and treatment decisions (higher burden). However, the majority of consumer-facing mental health AI tools fall outside FDA oversight entirely, occupying a regulatory gray area between wellness applications and medical devices (Grundy *et al*., 2019) [8].

The European Union's AI Act, passed in 2024, represents a more comprehensive regulatory framework, classifying AI systems used in healthcare as high-risk and imposing requirements for transparency, explainability, human oversight, and ongoing monitoring. Professional bodies including the American Psychological Association and the British Psychological Society have begun to develop guidance on the ethical use of AI-assisted tools in clinical practice, emphasizing that clinical responsibility cannot be delegated to algorithmic systems and that AI should function as a tool under human supervision rather than as an autonomous agent (APA, 2023) [2].

Toward a Human-Centered Integration Framework

The evidence reviewed above supports neither wholesale adoption nor categorical rejection of AI in mental health care. Rather, it points toward the need for a principled framework that captures the genuine potential of AI while mitigating its documented risks. We propose five pillars for such a framework.

First, clinical primacy: AI systems in mental health should function as decision-support tools under continuous clinician oversight, with no autonomous authority over diagnosis, treatment assignment, or risk intervention. The clinician-patient relationship, and the ethical responsibilities it entails, must remain the locus of clinical accountability.

Second, algorithmic transparency: mental health AI systems should be required to provide interpretable outputs that allow clinicians to understand and critically evaluate the basis of AI-generated assessments. "Black box" systems whose decision logic cannot be audited are incompatible with the standards of evidence-based practice.

Third, equity-centered design: AI systems should be developed and evaluated with explicit attention to performance across demographic subgroups, with mandatory disaggregated reporting of accuracy metrics. Training datasets should be curated to ensure adequate representation of historically underserved populations.

Fourth, robust data governance: mental health AI applications must be subject to stringent data minimization principles, meaningful consent processes, and prohibitions on commercial data sharing without explicit, informed authorization. Regulatory frameworks should close existing gaps that allow consumer wellness applications to collect sensitive mental health data without clinical oversight.

Fifth, rigorous evaluation: clinical deployment of AI tools should be preceded by robust randomized controlled trials conducted in representative populations, with outcomes that extend beyond short-term symptom reduction to include functional outcomes, therapeutic alliance, and long-term wellbeing.

Conclusion

The integration of artificial intelligence into mental health care represents both a historic opportunity and a significant ethical responsibility. The scope of the global mental health crisis is too vast, and the shortage of trained clinicians too severe, for the field to dismiss technologies that may extend the reach of effective care to underserved populations. At the same time, the vulnerabilities of individuals seeking mental health support — and the profound stakes of diagnostic and therapeutic error — demand that AI tools be held to rigorous standards of evidence, fairness, and accountability.

This paper has argued that the path forward is neither uncritical adoption nor reflexive rejection, but deliberate, equity-conscious integration grounded in human-centered values. Achieving this will require coordinated effort across research, clinical, regulatory, and commercial domains: researchers must generate more robust and representative evidence; clinicians must develop the competencies to critically evaluate AI tools; regulators must close existing oversight gaps; and technology developers must prioritize transparency and user wellbeing over engagement metrics.

Ultimately, AI in mental health should be understood not as a replacement for the therapeutic relationship — with all its irreducible human complexity — but as a potential amplifier of human care: extending its reach, sharpening its precision, and ensuring that its benefits are shared equitably across populations who have too long been left behind.

References

1. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M, *et al*. An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics,2020:132:103978. https://doi.org/10.1016/j.ijmedinf.2019.103978
2. American Psychological Association. Guidelines for the use of artificial intelligence in clinical psychology practice. APA Publications, 2023.
3. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M, *et al*. Ethical machine learning in healthcare. Annual Review of Biomedical Data Science,2021:4:123-144. https://doi.org/10.1146/annurev-biodatasci-092820-114757
4. Coppersmith G, Dredze M, Harman C, Hollingshead, K. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology, 2015, 1-10.
5. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF, *et al*. A review of depression and suicide risk assessment using speech analysis. Speech Communication,2015:71:10-49. https://doi.org/10.1016/j.specom.2015.03.004
6. Fitzpatrick KK, Darcy A, Vierhile M. 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,2017:4(2):19. https://doi.org/10.2196/mental.7785

7. Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJP, Dobson RJB, *et al*. Characterisation of mental health conditions in social media using informed deep learning. Scientific Reports,2017:7:45141. https://doi.org/10.1038/srep45141
8. Grundy Q, Chiu K, Held F, Continella A, Bero L, Holz R, *et al*. Data sharing practices of medicines-related apps and the mobile ecosystem. BMJ,2019:364:l920. https://doi.org/10.1136/bmj.l920
9. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR mHealth and uHealth,2018:6(11):12106. https://doi.org/10.2196/12106
10. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bulletin of the World Health Organization,2004:82(11):858-866.
11. Luxton DD. Artificial intelligence in psychological practice: Current and future applications and implications. Professional Psychology: Research and Practice,2014:45(5):332-339.
12. McCarthy JF, Bossarte RM, Katz IR, Thompson C, Kemp J, Hannemann CM, *et al*. Predictive modeling and concentration of the risk of suicide. Psychiatric Services,2015:66(11):1191-1198. https://doi.org/10.1176/appi.ps.201500014
13. Norcross JC, Wampold BE. Evidence-based therapy relationships: Research conclusions and clinical practices. Psychotherapy,2011:48(1):98-102.
14. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine,2016:375(13):1216-1219.
15. Obermeyer Z, Powers B, Vogeli C, Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science,2019:366(6464):447-453.
16. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli, A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease. Neuroscience & Biobehavioral Reviews,2012:36(4):1140-1152.
17. Patel V, Chisholm D, Parikh R, Charlson FJ, Degenhardt L, Dua T, *et al*. Addressing the burden of mental, neurological, and substance use disorders: Key messages from Disease Control Priorities, 3rd edition. The Lancet,2018:387(10028):1672-1685.
18. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, *et al*. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior. Journal of Medical Internet Research,2015:17(7):175. https://doi.org/10.2196/jmir.4273
19. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine,2019:49(9):1426-1448.
20. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, *et al*. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. American Journal of Psychiatry,2018:175(10):951-960.
21. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, *et al*. The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry,2021:20(3):318-335.
22. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science,2017:5(3):457-469.
23. World Health Organization. World mental health report: Transforming mental health for all. WHO Press, 2022.