The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 3 | Issue 7 | July 2025
ISSN: 2583-973X (Online)
Website: www.theacademic.in

The Transformative Role of Artificial Intelligence in Personalized Mental Health: A Comprehensive Analysis and Review

Abhishek Dey

Assistant Professor, Department of Computer Science, Bethune College, Kolkata, West Bengal, India

Email ID: dey.abhishek7@gmail.com

DOI : https://doi.org/10.5281/zenodo.16785309

ARTICLE DETAILS

Research Paper

Accepted: 18-07-2025

Published: 10-08-2025

Keywords:

Artificial Intelligence, Mental Health, Digital Phenotyping, Conversational Agents, Emotion Recognition, Explainable AI

ABSTRACT

The growing global burden of mental health disorders underscores the urgent need for innovative, scalable solutions that extend beyond conventional models of care. Artificial Intelligence (AI) has swiftly emerged as a transformative force, empowering the development of personalized approaches to mental health care and support. In this review, we have explored the current state and evolving landscape of AI applications in mental healthcare, focusing on key domains such as conversational agents, digital phenotyping, personalized intervention delivery, and affective computing. By analyzing recent advancements, we have highlighted how AI-powered systems have been designed to tailor mental health support through real-time emotion recognition, behavioral data analysis, and adaptive feedback. Empirical studies have been reviewed to assess the effectiveness and limitations of these technologies across diverse populations and clinical contexts. Technical foundations including machine learning architectures, natural language processing, and multimodal data integration have been discussed to elucidate the mechanisms behind these innovations. At the same time, we have addressed critical ethical and societal concerns, with particular attention to data privacy, algorithmic bias, and the delicate nature of human-AI interaction in emotionally sensitive scenarios. These issues must be thoughtfully navigated to ensure the safety and fairness of AI-driven interventions. Although AI holds significant promise, barriers such as lack of transparency, insufficient clinical validation, and regulatory uncertainty persist. To overcome these challenges, we have also proposed future research directions that emphasize the development of responsible and explainable AI, and robust evaluation frameworks.

Introduction

The escalating prevalence of mental health disorders and the growing recognition of the importance of holistic well-being underscore a critical global health challenge (Insel, 2018). Traditional mental healthcare systems often face significant barriers, including systemic stigma, limited access to qualified professionals, geographical disparities, and the inherent difficulty in delivering truly personalized care that accounts for individual variability in symptoms, responses to treatment, and socio-environmental factors (Naslund et al., 2019). Artificial Intelligence (AI) has emerged as a promising avenue to address these limitations by leveraging advanced computational power and data analysis to provide highly individualized support, continuous monitoring, and precision interventions.

This review paper has aimed to synthesize the current state of AI applications in personalized well-being and mental health. It has been explored how AI technologies are moving beyond generalized approaches to deliver tailored solutions, augmenting human capabilities rather than simply replacing them. The prominent AI methodologies have been categorized and discussed, their empirical efficacy has been analyzed, and the ethical, social, and practical implications of their deployment have been critically examined. Finally, key challenges have been outlined, and future research directions have been charted to guide responsible and impactful innovation in this burgeoning field.

2. AI Applications in Personalized Well-being and Mental Health

The integration of AI into mental health and well-being has manifested in several distinct yet interconnected application areas. Each area leverages specific AI techniques to deliver personalized support.

2.1. Conversational AI and Chatbots

AI-powered conversational agents (CAs), often deployed as chatbots or virtual assistants, have emerged as a highly accessible and potentially stigma-reducing means of delivering mental health support. These agents utilize sophisticated Natural Language Processing (NLP) and Natural Language Generation (NLG) techniques to engage users in human-like dialogue.

Mechanisms and Capabilities

Early chatbots, such as ELIZA, demonstrated the rudimentary ability to simulate therapeutic conversations (Weizenbaum, 1966). Modern AI-powered CAs leverage deep learning models, including Large Language Models (LLMs), to understand complex user input, recognize emotional states through sentiment analysis, and generate relevant, empathetic responses (Torous et al., 2021a). They often incorporate principles from evidence-based psychotherapies like Cognitive Behavioral Therapy (CBT) or Dialectical Behavior Therapy (DBT), guiding users through exercises, thought record completion, and coping strategies (Fitzpatrick et al., 2017; Fulmer et al., 2018). Examples include Woebot, Wysa, and Replika, which offer 24/7 availability, addressing a significant barrier to traditional care access.

Efficacy and Limitations

Recent systematic reviews and meta-analyses suggest that AI-driven CAs can have a moderate-to-large effect on improving depressive symptoms, particularly in subclinical populations (Feng et al., 2025). They are capable of providing psychoeducation, emotional support, and tools for self-management, proving beneficial for those seeking initial support or supplementary care. However, their efficacy for more severe conditions remains less clear, and concerns persist regarding their ability to genuinely understand complex emotional nuances, handle crisis situations, and form a therapeutic alliance comparable to human interaction (Inkster et al., 2018; Vaidyam et al., 2019).

2.2. Digital Phenotyping and Passive Sensing

Digital phenotyping refers to the continuous, real-time quantification of an individual’s behavioral patterns, emotional states, and mental health indicators using data derived from everyday interactions with personal digital devices such as smartphones, wearable sensors, and online platforms (Insel, 2017; Torous et al., 2020).

Data Sources and AI Techniques

AI plays a pivotal role in transforming the vast, heterogeneous, and often unstructured data generated by these sources into clinically meaningful insights. A range of data modalities are harnessed in this process. Smartphone sensors, including GPS, accelerometers, and screen usage logs, are analyzed through machine learning techniques such as Hidden Markov Models and Recurrent Neural Networks to identify anomalies in mobility, physical activity, sleep patterns, and social communication that may indicate shifts in mental state (Onnela & Rauch, 2016; Wang et al., 2020).

Wearable devices add physiological context by tracking heart rate variability, galvanic skin response, and sleep quality—biomarkers that have been linked to stress, anxiety, and depressive disorders (Picard et al., 2001). Voice and language data, captured through naturalistic sources such as phone conversations or voice notes, are analyzed using speech processing and natural language processing (NLP) techniques to detect vocal and linguistic biomarkers associated with emotional dysregulation and cognitive decline (Al Hanai et al., 2020; Cao et al., 2019).

Furthermore, public social media activity offers valuable insights into an individual's psychological state; through sentiment analysis and linguistic modeling of posts and interactions, AI can detect indicators of emotional distress, social isolation, or even thoughts of self-harm (De Choudhury et al., 2013; Reece & Danforth, 2017). Together, these multimodal digital phenotyping approaches, when combined with robust AI analytics, enable scalable, non-invasive, and context-aware mental health monitoring, potentially transforming early detection and personalized intervention in psychological care.

Applications

AI-enabled digital phenotyping supports a wide range of clinical and preventive applications. It allows for the early identification of emerging mental health issues, prediction of symptom exacerbation or relapse, and real-time tracking of treatment progress (Barnett et al., 2018; Jacobson et al., 2019). By continuously monitoring an individual's unique behavioral and physiological patterns, AI systems can deliver highly personalized alerts, interventions, and recommendations tailored to moment-to-moment fluctuations in psychological state. This not only improves the precision of mental health support but also enables proactive, just-in-time intervention—especially valuable in remote care or underserved populations.

Challenges

Despite its promise, digital phenotyping faces several technical, practical, and ethical challenges. The heterogeneity of data sources, along with the need for robust algorithms that generalize across diverse populations and contexts, presents a significant obstacle. In addition, high-frequency data collection raises substantial privacy concerns, particularly regarding informed consent, data ownership, and potential misuse in surveillance or profiling (Mohr et al., 2017). Ensuring fairness, transparency, and accountability in AI-driven phenotyping systems is essential for their ethical and equitable integration into mental healthcare.

2.3. Personalized Interventions and Treatment Optimization

AI is playing a pivotal role not only in the assessment and monitoring of mental health conditions but also in shaping personalized therapeutic strategies and optimizing treatment pathways. By leveraging advanced machine learning techniques—such as predictive modeling, reinforcement learning, and recommendation systems—AI enables more precise, adaptive, and individualized care. This shift marks a departure from traditional "one-size-fits-all" approaches toward a more nuanced, patient-centered paradigm.

Precision Psychiatry

AI-driven models have been increasingly applied in precision psychiatry, where they analyze large-scale, multimodal datasets—including electronic health records (EHRs), genetic information, neuroimaging data, and behavioral or digital phenotypes. These models aim to identify reliable biomarkers and predict patient-specific responses to various pharmacological and psychotherapeutic interventions (Dwyer et al., 2018; Rush et al., 2006). Such predictive capabilities hold the potential to drastically reduce the current reliance on trial-and-error treatment selection in clinical psychiatry, leading to improved clinical outcomes, reduced side effects, and faster recovery times.

Adaptive and Just-in-Time Interventions

AI technologies also enable the dynamic personalization of mental health interventions through real-time feedback mechanisms. Using reinforcement learning and decision optimization algorithms, digital platforms can adjust the type, intensity, and timing of interventions based on individual symptom trajectories, behavioral engagement, and contextual factors (Ben-Zeev et al., 2019). For instance, an AI system delivering cognitive behavioral therapy (CBT) modules may learn to prioritize specific therapeutic content—such as emotion regulation or cognitive restructuring—based on the user's evolving needs and interaction history. These adaptive interventions align closely with the "just-in-time adaptive interventions" (JITAIs) framework, which aims to deliver support precisely when individuals are most receptive or vulnerable.

Personalized Recommendation Systems

Drawing inspiration from recommender systems used in e-commerce and entertainment, AI is also being leveraged to deliver personalized suggestions for non-clinical well-being interventions. These systems can recommend mindfulness practices, stress-reduction techniques, physical activities, social engagement opportunities, or digital support communities tailored to the user's mental health profile, personal preferences, and behavioral patterns (Calvo & Peters, 2016). The aim is to maximize user engagement, adherence, and therapeutic efficacy while promoting autonomy and self-management in mental health care.

Integration into Clinical Workflows

Importantly, these AI-driven personalization approaches are increasingly being integrated into broader digital mental health platforms and clinical decision support systems. When used alongside human oversight, they enhance clinical efficiency, reduce clinician burden, and facilitate shared decision-making between patients and providers.

2.4. Affective Computing and Emotion Recognition

Affective computing represents a critical frontier in the development of emotionally intelligent artificial intelligence systems. It aims to enable machines not only to detect and interpret human emotions but also to respond in ways that are empathetic, contextually appropriate, and emotionally supportive. In the context of mental health care, the integration of affective computing is instrumental in enhancing personalization, emotional attunement, and user engagement, especially in applications like conversational agents, virtual therapists, and digital self-help tools.

Techniques and Modalities

Affective computing draws on a rich array of interdisciplinary techniques from computer vision, audio signal processing, natural language processing (NLP), and multimodal machine learning. Computer vision methods are used to analyze facial expressions, eye movements, microexpressions, and even posture or body language to infer affective states (Picard, 1997). Speech analysis leverages prosodic features such as pitch, tone, tempo, and vocal intensity to detect emotional nuances in spoken language (Scherer, 2003). Text-based NLP models apply sentiment analysis, emotion classification, and contextual understanding to written or transcribed speech, identifying emotions such as sadness, anger, fear, or joy. Multimodal deep learning models increasingly combine these inputs to improve accuracy, learning from large, annotated emotional datasets and user interactions. Recent advances in transformer-based models (e.g., BERT, GPT, wav2vec) and cross-modal attention mechanisms have further boosted emotion recognition capabilities, enabling more nuanced and real-time affect detection even in noisy, real-world environments.

Role in Personalization

Emotion recognition plays a pivotal role in enhancing the responsiveness and relevance of AI-powered mental health interventions. By accurately identifying a user’s emotional state in real time, AI systems can tailor their responses and actions accordingly. For example, a digital mental health assistant may detect signs of frustration in a user’s tone and shift its language to be more supportive and calming. If high stress or anxiety is detected, a personalized wellness platform could proactively suggest relaxation exercises, breathing techniques, or mood-lifting activities. Conversational agents might adjust pacing, topic shifts, or even escalate the interaction to a human clinician if signs of distress or suicidality are inferred. This capacity for emotional attunement significantly enhances user trust, perceived empathy, and adherence to digital interventions—key factors for effective engagement in mental health support.

Challenges and Limitations

Despite rapid progress, affective computing faces significant challenges, especially when deployed in sensitive contexts like mental health care.

Emotion Complexity

Emotions are complex, dynamic, and often ambiguous. A single expression or tone may reflect multiple, overlapping emotional states.

Context Dependency

Emotional cues are highly context-dependent; the same expression or phrase may convey different emotions based on situational, cultural, or individual factors.

Cultural and Demographic Bias

Most affective datasets are skewed toward Western or adult populations, raising concerns about cultural insensitivity and demographic bias in emotion recognition algorithms (Barrett et al., 2019).

Privacy and Ethics

The inference of emotions from user data, especially covertly or without explicit consent raises ethical concerns about surveillance, emotional manipulation, and autonomy.

Addressing these limitations requires interdisciplinary collaboration among AI researchers, psychologists, linguists, and ethicists to develop models that are not only technically robust but also socially responsible and culturally inclusive.

3. Ethical Considerations and Challenges

The integration of AI into personalized mental health care presents vast opportunities, yet it is accompanied by a complex array of ethical dilemmas and implementation challenges. As AI systems increasingly influence how mental health is assessed, monitored, and treated, safeguarding individual rights, promoting equitable access, and ensuring trustworthiness become paramount. This section examines the core ethical and practical concerns that must be addressed to ensure the responsible and sustainable deployment of AI in mental health contexts.

3.1. Data Privacy, Security, and Ownership

Mental health-related data is arguably among the most sensitive types of personal information, encompassing private details about an individual's emotional states, behavioral patterns, and therapeutic histories. The continuous and often passive collection of such data—especially through smartphones, wearables, and online interactions—raises critical concerns regarding privacy and security.

Highly Sensitive Data

Data generated from AI-driven mental health tools may include real-time mood tracking, voice notes expressing distress, or indicators of suicidal ideation. The misuse or leakage of such data could have devastating consequences for individuals. Robust data encryption, secure storage architectures, anonymization protocols, and multi-level access control mechanisms are essential to protect against breaches and unauthorized use (Luxton, 2014).

Informed Consent

Achieving meaningful informed consent in digital mental health interventions is especially challenging. Users—particularly those in psychological distress—may not fully grasp the extent of data being collected, how it is processed by opaque algorithms, or who ultimately has access. Continuous, contextual, and dynamic consent mechanisms are needed to improve comprehension and autonomy (Nebeker et al., 2019).

Data Ownership and Control

Clear policies must define who owns the data generated by AI mental health tools. Users should have transparent access to their data, the ability to delete or modify it, and control over third-party data sharing. The current lack of global data governance frameworks complicates this issue, particularly across cross-border applications and cloud-based platforms.

3.2. Algorithmic Bias and Fairness

AI systems inherit the biases of the data they are trained on, which can lead to inequitable mental health outcomes if not properly addressed. These biases can manifest across several dimensions—from representational disparities to culturally inappropriate interpretations.

Representational Bias

Many AI models in mental health are trained on datasets that overrepresent certain populations—typically young, urban, tech-literate, and Western individuals. This leads to poor generalization to underrepresented groups such as ethnic minorities, older adults, or individuals in low-resource settings, resulting in reduced diagnostic accuracy and exclusion from effective care (Ghassemi et al., 2018).

Diagnostic and Treatment Bias

Algorithmic decision-making may lead to unequal treatment recommendations or misdiagnoses based on race, gender, socioeconomic background, or disability status. For example, AI systems might overdiagnose certain disorders in one demographic while under-diagnosing in another, exacerbating existing healthcare inequalities (Chen et al., 2021).

Cultural Sensitivity

AI systems often lack cultural competence. Mental health symptoms and expressions vary greatly across cultures—what constitutes depression or anxiety in one society may not be recognized as such in another. Without explicit efforts to incorporate cultural variation into training data and system design, AI risks promoting a homogenized, Western-centric view of mental health (Torous & Roberts, 2018).

3.3. Trust, Empathy, and the Therapeutic Alliance

While AI enables scalable and immediate mental health support, it cannot yet replicate the complex emotional intelligence, intuition, and interpersonal warmth that human therapists provide. This raises questions about the depth and authenticity of human-AI interaction in therapeutic settings.

Human Connection

The therapeutic alliance—characterized by empathy, trust, and attunement—is foundational to effective mental health treatment. While AI can simulate supportive dialogue, it currently lacks the capacity for true emotional understanding or ethical judgment. Users may sense this limitation, potentially undermining therapeutic engagement (Norcross & Lambert, 2011; Minor et al., 2020).

Over-reliance and Dehumanization

There is a growing concern that an over-dependence on AI systems could lead to the dehumanization of care. Individuals may begin to see digital interventions as replacements rather than supplements to human care, especially in contexts where mental health professionals are scarce. This is particularly dangerous in high-risk situations—such as suicidal ideation—where nuanced human judgment is irreplaceable.

Ethical Black Box

Many high-performing AI models, especially deep neural networks, function as “black boxes”—they offer accurate predictions or recommendations without clear explanations of how decisions are made. This opacity complicates ethical accountability, limits clinicians’ ability to critically assess AI output, and hinders user trust in the technology (Adadi & Berrada, 2018).

3.4. Regulatory Frameworks and Clinical Validation

The pace of innovation in AI-enabled mental health care has outstripped the development of rigorous regulatory frameworks and evidence-based validation protocols. Without proper oversight, the risk of harm from inaccurate or untested technologies increases.

Lack of Standards

Regulatory frameworks for AI in mental health are currently disjointed and vary significantly across jurisdictions. There is a pressing need for regulatory bodies to establish clear guidelines on data governance, clinical validation, algorithm transparency, and ethical standards for digital mental health tools (Hollis et al., 2018).

Clinical Evidence Gap

While many AI interventions show promise in feasibility studies, few have been subjected to rigorous randomized controlled trials (RCTs) to establish their clinical efficacy and safety at scale. Long-term studies are needed to assess real-world impact, unintended consequences, and outcomes across diverse populations (Neary & Schueller, 2018).

Risk of Harm and Misinformation

Without adequate safeguards, AI systems may deliver inaccurate diagnoses or harmful advice, especially in cases involving self-harm or suicidal behavior. Misinterpretation of user input or system failure in crisis scenarios could have life-threatening implications. Ensuring continuous human oversight and fail-safes is critical (Oh et al., 2019).

4. Future Directions and Research Opportunities

The continued advancement of AI in personalized mental health care hinges not only on technological innovation but also on addressing unresolved challenges in ethics, usability, and clinical integration. To fully realize AI's transformative potential, research must shift toward building safe, explainable, and inclusive systems that enhance human capabilities, respect user agency, and promote long-term well-being. This section outlines five key directions that represent both opportunities and imperatives for future work in this domain.

4.1. Hybrid Human-AI Models and Augmented Intelligence

The future of mental health AI lies in the augmentation—not replacement—of human clinicians. Hybrid human-AI frameworks emphasize collaboration, where AI acts as an intelligent assistant, supporting decision-making and improving access without displacing the uniquely human aspects of therapy.

AI as a Clinical Assistant

AI systems can synthesize complex patient data, monitor mental health trends over time, identify risk markers, and generate personalized intervention recommendations. These tools can alleviate the cognitive load on clinicians and enable more informed, data-driven decisions, particularly in high-volume or resource-limited settings (Torous et al., 2019).

Seamless Integration

Future systems should focus on embedding AI functionalities into clinical workflows and electronic health records through intuitive, user-friendly interfaces. The goal is to create an unobtrusive, supportive tool that enhances efficiency without disrupting therapeutic continuity.

Telehealth Enhancement

AI can play a critical role in enhancing telepsychiatry services by enabling passive monitoring, automated appointment reminders, personalized self-help content, and just-in-time interventions. These enhancements help bridge gaps in care, especially in rural or underserved areas where mental health professionals are scarce (Luxton et al., 2020).

4.2. Explainable AI (XAI) for Transparency and Trust

In high-stakes domains like mental healthcare, trust is foundational. Yet, many powerful AI models operate as opaque "black boxes," leaving both users and clinicians uncertain about how decisions are made. Explainable AI (XAI) research is therefore critical.

Interpretability and Accountability

Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) allow stakeholders to understand which features most influenced an AI's output (Lundberg & Lee, 2017; Ribeiro et al., 2016). These tools make AI more transparent, support auditing and accountability, and empower clinicians to challenge or override model outputs when necessary.

Clinical Decision Support

XAI enables clinicians to interpret AI-generated insights within the broader therapeutic context, improving their ability to make balanced, personalized decisions and reducing the risk of over-reliance on automated tools.

User Empowerment

Providing users with comprehensible explanations enhances trust, supports autonomy, and improves engagement. For example, showing users how certain behaviors influenced a stress score can increase awareness and adherence to interventions.

4.3. Multimodal Data Fusion and Advanced AI Architectures

To move beyond surface-level insights, future AI systems must be capable of understanding the complexity of human emotion, behavior, and cognition by integrating diverse data sources through advanced architectures.

Beyond Unimodal Analysis

Mental health is inherently multimodal, involving signals from voice, text, facial expression, physiological data, and contextual information. Future systems should combine these inputs using attention-based fusion architectures and graph neural networks to derive a holistic, nuanced understanding of mental states (Ringwald et al., 2021).

Personalization through Reinforcement Learning

Reinforcement learning algorithms allow systems to continuously learn from user feedback, adapting content and delivery in real time. These agents can personalize intervention timing, modality, and complexity based on engagement patterns and evolving needs (Ghassemi et al., 2020).

Federated Learning for Privacy

Federated learning allows AI models to be trained across decentralized data sources—such as user smartphones or clinical sites—without transferring sensitive data to central servers. This approach enhances both personalization and privacy while addressing regulatory constraints (Rieke et al., 2020).

4.4. Proactive and Preventative Mental Health

Shifting from reactive treatment to proactive and preventative mental healthcare is a promising future direction. AI can be instrumental in identifying at-risk individuals early, promoting mental resilience, and supporting mental well-being throughout the lifespan.

Early Risk Detection

AI algorithms can flag early indicators of mental health deterioration based on subtle changes in behavior, speech, or physiological data—enabling timely, low-intensity interventions before symptoms escalate (Chandrashekar, 2020).

Personalized Resilience Building

Future tools can move beyond diagnosis and treatment by recommending personalized exercises in mindfulness, emotion regulation, cognitive reappraisal, and social connection tailored to the individual's profile and preferences (Calvo & Peters, 2016).

Population-Level Insights

AI can analyze large-scale, de-identified datasets to uncover emerging mental health trends, assess the effectiveness of public interventions, and inform targeted resource allocation. These insights can guide governments and NGOs in designing population-level mental health strategies.

4.5. Ethical AI by Design and Policy Development

For AI to become a trusted ally in mental health, ethical principles must be embedded into every stage of its development and deployment. Equally important are policy frameworks that enable innovation while safeguarding human rights.

Privacy-Preserving AI

Techniques such as differential privacy, homomorphic encryption, and secure multi-party computation can protect individual data while still allowing for powerful analytics (Dwork et al., 2006). These methods must become standard in mental health AI research.

Bias Mitigation Pipelines

Developing systematic processes for identifying and correcting bias—across data collection, model training, and deployment—is crucial. These pipelines must include fairness metrics, representational audits, and adversarial testing (Mehrabi et al., 2021).

Regulatory Sandboxes and Guidelines

Governments and health agencies should establish regulatory sandboxes that allow controlled testing of novel AI applications under supervision. Simultaneously, the development of standardized guidelines for safety, efficacy, transparency, and certification is essential for fostering responsible innovation (European Commission, 2020).

Public Engagement and Literacy

Promoting AI and mental health literacy is essential for enabling informed participation. Public campaigns, education in schools and universities, and transparent communication about risks and benefits can empower individuals and reduce fear or misinformation.

5. Conclusion

Artificial Intelligence has shown immense promise for revolutionizing personalized well-being and mental health by providing scalable, accessible, and highly individualized support. From the empathetic dialogue of conversational agents to the subtle insights gleaned from digital phenotyping and the precision of personalized interventions, AI is poised to fundamentally reshape how individuals manage their mental health. However, this transformative potential is intrinsically linked to the responsible navigation of profound ethical challenges, particularly concerning data privacy, algorithmic bias, and the imperative to preserve the deeply human element of mental health support. The future of AI in this domain is envisioned not as one of human replacement but of human augmentation. By prioritizing the development of explainable, trustworthy, and culturally sensitive AI solutions, fostering hybrid human-AI collaborative models, and establishing robust regulatory and ethical frameworks, AI can be harnessed to create a more accessible, equitable, and effective ecosystem for personalized well-being and mental health globally. Continued interdisciplinary collaboration among AI researchers, mental health professionals, ethicists, policymakers, and user communities will be critical to ensuring that AI serves humanity's well-being in a meaningful and ethical way.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.
- Al Hanai, T., Maes, P., & Picard, R. W. (2020). Using AI to monitor speech for mental health. *The American Journal of Psychiatry*, *177*(9), 834–836.
- Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., & Onnela, J. P. (2018). Relapse prediction in schizophrenia through digital phenotyping: A pilot study. *Neuropsychopharmacology*, *43*(8), 1660–1666.
- Barrett, L. F., Adolphs, R., Marsella, S., & Stern, A. (2019). The human emotional experience is not what you think. *Trends in Cognitive Sciences*, *23*(10), 785–796.
- Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2019). Next-generation psychiatric mobile health: Personalized interventions via machine learning. *Psychiatric Rehabilitation Journal*, *42*(3), 221–229.
- Calvo, R. A., & Peters, D. (2016). *Positive computing: Technology and well-being*. MIT Press.
- Cao, R., Ni, J., & Cai, B. (2019). *Speech emotion recognition with feature enhancement and multi-scale fusion network*. Paper presented at the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK.
- Chandrashekar, P. (2020). Artificial intelligence in mental healthcare: A review of technological advancements and ethical issues. *Current Opinion in Psychiatry*, *33*(6), 578–583.
- Chen, I., Szolovits, P., & Ghassemi, M. (2021). Can AI be fair and trustworthy in medical diagnosis? *Artificial Intelligence in Medicine*, *111*, 101991.
- De Choudhury, M., Counts, S., & Horvitz, E. (2013). *Social media as a measurement of mood and wellbeing*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Paris, France.
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. P. Vadhan (Ed.), *Theory of cryptography conference* (pp. 265–284). Springer.
- Dwyer, D. B., Erus, G., & Satterthwaite, T. D. (2018). Prediction of psychiatric illness from neuroimaging: Promising results and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(11), 1083–1090.

- European Commission. (2020). *White Paper on Artificial Intelligence – A European approach to excellence and trust* (COM(2020) 65 final).
- Feng, Y., Hang, Y., Wu, W., Song, X., Xiao, X., Dong, F., & Qiao, Z. (2025). Effectiveness of AI-driven conversational agents in improving mental health among young people: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *27*(1), e69639.
- Fitzpatrick, K. K., 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 Mental Health*, *4*(2), e19.
- Fulmer, R., Joerin, A., Gentile, A., & Sequeira, L. (2018). Using psychotherapeutic artificial intelligence to address adolescent mental health. In *Proceedings of the International Conference on Internet Science* (pp. 58–72). Springer.
- Ghassemi, M., Celi, L. A., & Stone, P. (2020). The false hope of current approaches to AI in healthcare. *The Lancet Digital Health*, *2*(9), e484–e487.
- Ghassemi, M., Naeini, M. P., & Naeini, R. (2018). *Fairness in machine learning: A survey*. arXiv preprint arXiv:1808.00684.
- Hollis, C., Salisbury, H., Tang, V., Thake, A., ... & King, H. (2018). Digital technologies for the management of mental health problems in children and young people: A systematic review. *Journal of Child Psychology and Psychiatry*, *59*(5), 476–500.
- Inkster, B., Sarda, S., & Subramanian, V. (2018). An artificial intelligence-enabled chatbot for mental health support: A study of content analysis and user engagement. *Journal of Medical Internet Research*, *20*(5), e10403.
- Insel, T. R. (2017). Digital phenotyping: A new tool for mental health research. *JAMA*, *318*(13), 1215–1216.
- Insel, T. R. (2018). The promise of digital psychiatry: A new dataset for mental health research. *World Psychiatry*, *17*(1), 75–76.
- Jacobson, N. C., Chung, J., & Tong, D. C. (2019). Machine learning for predicting psychosis: A systematic review. *Schizophrenia Research*, *216*, 2–10.
- Luxton, D. D. (2014). Artificial intelligence in psychological practice: Current and future applications and ethical considerations. *Professional Psychology: Research and Practice*, *45*(5), 332–338.

- Luxton, D. D., Torous, J., & Wykes, T. (2020). Digital mental health and COVID-19: Opportunities and challenges. *JMIR Mental Health*, *7*(1), e19543.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(3), 1–35.
- Minor, B., Majumder, S., & De Choudhury, M. (2020). *Human-AI teaming for mental wellbeing: Opportunities and challenges*. Paper presented at the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA.
- Mohr, D. C., Lyon, A. R., & Hunter, C. M. (2017). Digital phenotyping and the future of mental health care. *JAMA*, *318*(13), 1217–1218.
- Naslund, J. A., Marsch, L. A., & Greden, J. F. (2019). Digital mental health: Opportunities and challenges. *World Psychiatry*, *18*(2), 159–160.
- Neary, M., & Schueller, S. M. (2018). The current state of mobile applications for depression: A narrative review of efficacy, engagement, and usability. *Clinical Psychology Review*, *64*, 1–18.
- Nebeker, C., Harlow, J., & Prochaska, J. J. (2019). Ethical concerns with emerging digital technologies in behavioral health. *Journal of Behavioral Health Services & Research*, *46*(2), 291–301.
- Norcross, J. C., & Lambert, M. J. (2011). Psychotherapy relationships that work. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Evidence-based responsiveness* (2nd ed., pp. 3–25). Oxford University Press.
- Oh, S. T., Kim, M. A., & Cha, H. (2019). Ethical challenges of artificial intelligence in healthcare. *Journal of the Korean Medical Association*, *62*(11), 598–604.
- Onnela, J. P., & Rauch, S. L. (2016). Harnessing information technology for behavioral health. *New England Journal of Medicine*, *374*(9), 805–806.
- Picard, R. W. (1997). *Affective computing*. MIT Press.
- Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191.

- Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1), 15.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the predictions of any classifier*. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA.
- Rieke, N., Hancox, J., Li, W., & Others. (2020). The future of digital health with federated learning. *npj Digital Medicine*, *3*(1), 119.
- Ringwald, C., O'Connor, K., & Hanrahan, J. (2021). Multimodal fusion for mental health diagnosis: A systematic review. *JMIR Medical Informatics*, *9*(3), e24945.
- Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., Niederehe, G., Thase, M. E., Lavori, P. W., Lebowitz, B. D., McGrath, P. J., Rosenbaum, J. F., Sackeim, H. A., Kupfer, D. J., Luther, J., & Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry*, *163*(11), 1905–1917.
- Scherer, K. R. (2003). Psychological models of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 137–162). Oxford University Press.
- Torous, J., Bucci, S., Bell, I. H., Kessing, L. L. V., Faurholt-Jepsen, M., Whelan, R., Arenella, L., Baker, L. T., Bertagnoli, A., Bondar, N. D., Bosc, P., Bowditch, E. S., Brockman, E. T., Caton, H., Centeno, V., Christensen, H., Cohn, R. A., Cohen, D. M., Cook, M. L., ... & Christensen, H. (2021a). The growing field of digital psychiatry: Current evidence and the future of clinical care. *World Psychiatry*, *20*(2), 221–230.
- Torous, J., Hsin, H., & Greden, J. F. (2019). The role of artificial intelligence in mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Journal of Psychiatric Practice*, *25*(1), 1–13.
- Torous, J., Kiang, M. V., Lorme, J., & Onnela, J. P. (2020). Digital phenotyping for mental health: A systematic review of the state of the art. *Current Psychiatry Reports*, *22*(4), 16.
- Torous, J., & Roberts, L. W. (2018). The ethical implications of digital mental health technologies. *Journal of Nervous & Mental Disease*, *206*(8), 603–605.
- Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Hanauer, D. A., & Rodebaugh, T. L. (2019). Chatbots for the delivery of mental health interventions: Systematic review and meta-analysis. *JMIR Mental Health*, *6*(10), e16823.

- Wang, R., Scherer, E. A., & Campbell, A. T. (2020). Wearable sensing for mental health applications: A comprehensive review. *Journal of Biomedical and Health Informatics*, *24*(10), 2824–2838.
- 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.