



September 2026
ISSN:1538-1080

Volume 26 Number 3
<https://doi.10.78717/ijhc.2026.3.7>

ETHICS

Selected Ethical Considerations of AI Therapy **Jan Warner, PhD, LISW-S**

Abstract

This article examines selected ethical issues in AI-mediated psychotherapy, focusing on measuring its clinical efficacy and noting the limited evidence available for specific psychological disorders and populations. Current AI therapy research often relies on established measures of symptom reductions that were not designed specifically to measure AI psychotherapy and may overlook characteristics unique to AI-mediated treatment. This lack of validated AI-specific symptom outcome measures complicates the evaluation of research and commercial claims. Additional concerns include inconsistent definitions of AI interventions, limited long-term follow-up measures, cultural generalizability, digital literacy, underrepresentation of diagnostic and demographic groups, and the potential role of AI in assessment and diagnosis. Greater interdisciplinary collaboration, conceptual precision, culturally responsive research, standardization of terminology, validated methods of evaluating therapeutic processes and clinical outcomes are needed before AI interventions can be considered an ethical mental health care practice.

Keywords: artificial intelligence, psychotherapy, ethics, mental health, AI therapy

Selected Ethical Considerations of AI Therapy

The continued advancement of artificial intelligence (AI) has created new potential for delivering mental health care while also creating clinical, regulatory, methodological, and ethical challenges. Ethical challenges include determining the efficacy of AI-mediated interventions and establishing their appropriateness for specific psychological disorders and populations. In addition, how to adapt longstanding traditional ethical principles such as confidentiality, informed consent, privacy, professional competence, and professional responsibility within the technological paradigm has not been established. While complex, addressing these issues remains essential before AI psychotherapy can be considered an ethical part of mental health care.

Efficacy

One of the most significant ethical challenges surrounding AI-assisted psychotherapy concerns the evaluation of its clinical effectiveness. Ethical implementation presupposes that intervention is both safe and beneficial; however, the methods used to assess the effectiveness of AI psychotherapy remain underdeveloped. Numerous studies have evaluated AI interventions using established psychological outcome measures primarily for mood disorders. In fact, for AI healthcare, “A lack of standardized evaluation approaches has led to diverse and inconsistent methods, making comparing chatbot performance difficult. Moreover, there is no widely accepted guideline or framework for evaluating health care chatbots” (Hua et al., 2025, p. 2).

Most existing research outcome measures rely on clinical instruments that measure interventions with a human therapist. “The most commonly used scales

to measure depressive and anxiety symptoms were the Patient Health Questionnaire (PHQ; PHQ-9 and PHQ-8; 13/21, 62%) and the Generalized Anxiety Disorder-7 (GAD-7; 10/21, 48%), respectively” (Villarreal-Zegarra et al., 2024, Results section, para. 2). These measures possess strong psychometric properties and have demonstrated validity for assessing symptom severity and treatment outcomes within traditional human-delivered psychotherapy. Nevertheless, they were not developed to evaluate AI-mediated therapeutic interactions. Consequently, although these instruments provide valuable information regarding psychological symptoms, they may fail to capture dimensions unique to AI-therapy. Those dimensions may be clinically useful, may lack value, or be harmful to clients.

Future evaluation of AI psychotherapy will retain many traditional measures of therapeutic success. However, AI-mediated psychotherapy is also likely to require additional constructs that reflect characteristics unique to interactions with artificial intelligence. Giacobello (2025) emphasized the importance of explicability in AI therapy, stating, “Explicability, requiring that AI processes be comprehensible and transparent, is intrinsically linked to the principle of autonomy and its practical expression: informed consent” (p. 1). Such factors may become essential components of future outcome measures specifically designed for AI therapy. At present, the Digital Working Alliance Inventory (DWAi) is a widely known validated instrument designed to assess the therapeutic relationship within digital mental health interventions (Goldberg et al., 2022). Developed for use with unguided digital mental health applications, the DWAi specifically only assesses users' perceptions of the working

alliance with a digital intervention. Goldberg et al. (2022) evaluated the DWAI as a measure of digital therapeutic alliance based on Bordin's (1979) concepts of therapeutic goals, tasks, and bond, finding evidence supporting its reliability and validity in unguided digital interventions. This may call into question the longstanding assumption that a therapeutic alliance necessarily requires a human therapist and suggest that meaningful therapeutic relationships may develop in digital mental health settings. Nevertheless, additional research is needed to determine the extent to which these digital alliances contribute to clinically meaningful improvements in mental health outcomes.

Despite its strengths, the DWAI is not designed to measure treatment effectiveness. It does not directly measure symptom reduction, improvements in psychosocial functioning, treatment response, or long-term recovery, nor was it developed to compare differences between digital interventions and therapist-delivered psychotherapy. Consequently, an important gap remains in the evaluation of AI psychotherapy: validated measures specifically designed to evaluate the clinical effectiveness of AI therapy have yet to be developed. Currently, researchers employ various instruments. For example, Jabir et al. (2023) reviewed AI therapy studies and reported 203 different outcome measurement instruments; and even more unfortunately, validity evidence was not cited for more than half of these instruments. Until such standard measures become available, claims regarding the efficacy of AI psychotherapy should be interpreted with appropriate ethical caution.

An additional ethical challenge concerns the disciplinary differences that shape how AI interventions are evaluated. Cho et al. (2023) observed that computer science research typically emphasizes

technical performance indicators, including user response quality, algorithmic accuracy, and system functionality. In contrast, medical and psychological research focuses primarily on clinical outcomes, such as symptom reduction, treatment adherence, therapeutic alliance, and improvements in functioning. These differing evaluation frameworks employ distinct terminology, methodologies, and outcome measures, making comparisons across disciplines difficult and limiting the development of a unified understanding of AI's ethical determinants including therapeutic effectiveness. Greater interdisciplinary collaboration between mental health professionals and computer scientists will likely be essential for establishing standardized methods for evaluating AI psychotherapy.

The current evidence base is further limited by several shortcomings. From an ethical perspective, the absence of long-term follow up measures of symptoms creates a particular problem: short-term symptom improvement cannot by itself establish that a therapeutic intervention produces durable benefit. This restricts understanding of the durability of treatment effects over time. Questions also remain regarding the generalizability of existing outcomes across cultural contexts. Barnes et al. (2024) reported that their "review suggests that cultural identity shapes how individuals include AI in constructing the self in relation to others and determines the effect of AI on key decision-making processes" (p. 2). Questions also remain regarding the generalizability of existing findings across cultural contexts. Much of the psychological research literature has historically relied on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations, raising concerns about whether findings generalize across cultures (Henrich et al., 2010; Tindle, 2021). This issue may be particularly

important for AI psychotherapy because cultural factors can influence how individuals perceive and interact with AI.

Investigators are beginning to look at issues with cultural differences between the AI therapy system designers and the users themselves (Barnes et al., 2024). Barnes et al. (2024) reviewed evidence suggesting that cultural identity, such as individualism and collectivism, influences how individuals perceive and interact with AI. Barnes et al. (2024) also identified social concepts about uncertainty avoidance and power distance as culturally shaped dimensions that may be relevant to how individuals conceptualize and respond to AI.

Research suggests that patterns of online communication may differ substantially across user groups. Kang and Shin (2016), for example, reported that individuals who are frequent mobile phone users tend to disclose more personal information when interacting with non-personalized digital sites—an especially noteworthy finding when the information being disclosed may be deeply personal or emotionally sensitive. Indeed, this increased willingness to self-disclose may facilitate therapeutic engagement within AI-mediated interventions.

Characteristics of individual users, including differences in digital literacy and technological competence, may substantially influence both therapeutic engagement and treatment outcomes. Yin et al. (2024) identified limited digital literacy—particularly among older adults—as a significant barrier to successful engagement with digital mental health services. Individuals who struggle to navigate digital technologies may experience reduced therapeutic benefit regardless of the quality of the AI intervention itself. These findings suggest that digital literacy should be considered an important moderating variable in future evaluations of AI psychotherapy.

Cultural factors present a related concern. Barnes et al. (2024) suggest that AI systems should be designed and evaluated with cultural context in mind rather than assuming findings from one culture generalize globally. Extending this work, Wang et al. (2026) conducted a meta-analysis on the moderating effect of national culture on AI adoption. They reviewed 85 independent samples across 25 countries. Their results provide quantitative evidence that AI adoption models should account for cultural differences instead of assuming universal patterns.

Lack of Research on Use with Specific Disorders and Populations

Another major ethical concern is the limited representation of psychological disorders and populations in AI psychotherapy research, with many disorders receiving little empirical attention and leaving substantial gaps in the evidence base. Therefore, for ethical purposes, the role of AI for these disorders must be approached cautiously. Although common mental disorders are recognized within contemporary psychological classification systems, AI psychotherapy research remains highly uneven across diagnostic populations. Research on AI in mental health care extends across a range of disorders, including schizophrenia; however, research specifically examining AI-delivered psychotherapy has concentrated more heavily on depression and anxiety. A review of the targeted outcomes reported by Zhang et al. (2025) indicates that personality disorders, hoarding disorder, body dysmorphic disorder, trichotillomania, and excoriation disorder were not represented among the 26 included studies. Consequently, broad claims regarding the effectiveness of AI psychotherapy across the full spectrum of mental illness are currently

unsupported and would be ethically difficult to justify.

The limited evidence base also raises important questions regarding the ethical use of AI in psychological assessment and diagnosis. Although standardized diagnostic systems such as the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2022) and insurance reimbursement frameworks rely on categorical diagnoses, clinical diagnosis frequently involves nuanced professional judgment that extends beyond the application of diagnostic criteria alone. Diagnostic decisions may have significant legal, occupational, educational, and financial consequences for individuals. In reviewing the literature for this article, additional ethical concerns were identified for AI diagnosing conditions involving substance use disorders, suicidality, cognitive impairment, or severe mental illness, particularly regarding fitness to drive, employment eligibility, legal proceedings, parenting abilities, disability income, and access to services. Consequently, whether AI systems should participate in diagnostic decision-making for psychological disorders remains an important ethical question requiring considerably more empirical investigation before widespread clinical implementation can be ethically recommended.

A further ethical challenge concerns the inconsistent definition of what constitutes an "AI therapy type intervention." The term is frequently applied to markedly different technologies, making comparisons across studies difficult. Early on, Weizenbaum (1966) explored using ELIZA—a computer program for the study of natural language communication between man and machine. This was an important historical precursor to modern therapeutic chatbots. Weizenbaum (1966) reported it was designed to simulate Rogerian-style

therapeutic conversation, a popular type of therapy at the time. Currently, cognitive behavioral therapy (CBT) is the most commonly used therapeutic orientation in AI-based mental-health chatbots (Kuhail et al., 2025).

However, sometimes CBT is not used, and no therapy is used. Yang et al. (2026) identified 83 studies involving AI in schizophrenia rehabilitation management, but most focused on symptom monitoring, medication management, or risk management; three examined psychosocial support. Notably, psychotherapy was not identified as a category of intervention in the 83 study review. Consequently, although these studies contribute valuable knowledge regarding AI in some mental health care, they provide limited evidence regarding the effectiveness of AI in other populations, other diagnoses, or psychotherapy.

A few commercial programs explicitly present themselves as therapy. One example is AI-Therapy, a web-based social anxiety software package whose current website describes 'Overcome Social Anxiety' as an evidence-based cognitive behavioral therapy program (AICBT Ltd., n.d.). It presents its 2018 randomized controlled trial as producing "very strong" results, including claims that its effect size was triple that of other standalone treatments and that the program was comparably effective to therapist-assisted interventions (AICBT Ltd., n.d.). The peer-reviewed study itself should be distinguished from these commercial claims. McCall et al. (2018) evaluated 'Overcome Social Anxiety' as a self-directed, web-based CBT program consisting of seven modules, individualized content based on users' questionnaire responses, automated email reminders, and prerecorded material from psychologists. It was not a modern generative-AI or large-language-model conversational therapist.

McCall et al. (2018) explicitly described the participants in the study regarding the website as a nonclinical sample of university students. This refers to their not being evaluated or diagnosed with any type of anxiety disorder. Participants reported social anxiety symptoms. The authors also disclosed that Fjola, D. Helgadóttir, the third author to the study, was a director of the company that created 'Overcome Social Anxiety,' and that, as a psychologist, she provided critical input into the study design. The paper's Conflicts of Interest section disclosed that Fjola, D. Helgadóttir was a company director and states that she did not fund the study and did not have access to the data. The other authors declared no conflicts of interest. Accordingly, the positive findings provide evidence concerning this particular web-based CBT intervention, population, outcomes, and study conditions. The findings should not be generalized as establishing effectiveness for clinically diagnosed populations or as evidence supporting AI psychotherapy overall. The distinction between the commercial name AI-Therapy, the technology actually evaluated, the population studied, and the scope of the supporting evidence illustrates the need for precision when evaluating commercial AI-psychotherapy claims and evaluating ethical concerns.

Beyond ambiguity regarding the therapy offered, the absence of standardized terminology also contributes to conceptual ambiguity within the literature. Terms such as large language model, AI-assisted therapy, AI-supported therapy, AI-enabled mental health care, digital mental health, technology-assisted psychotherapy, therapeutic chatbot, AI coach, and AI companion are frequently used interchangeably. Yet these differ substantially in their purpose, level of autonomy, and clinical scope. Some systems

provide structured psychotherapy, whereas others offer coaching, psychoeducation and so on. Without greater conceptual precision, meaningful comparisons between studies become increasingly difficult, and clinicians may struggle to determine whether published findings are applicable to the intervention being evaluated. As Kolding et al. (2025) concluded, "the field of generative AI in psychiatry and mental health is in its infancy, though evolving and growing exponentially" (p. 12).

Ultimately, the ethical acceptability of AI psychotherapy should depend not on the novelty or accessibility of the technology, but on whether its clinical benefits, limitations, and risks can be demonstrated with sufficient rigor. The current literature suggests that AI is already expanding into mental health care, yet unresolved ethical questions concerning its efficacy, diagnostic appropriateness, cultural generalizability, digital literacy, terminology, and the validity of commercial claims make broad clinical adoption premature. In addition, many other ethical issues remain unresolved that are beyond this article's scope, such as clinical, regulatory, methodological issues and transparency regarding digital storage and digital sharing of personal health information.

A defensible path forward therefore requires more than adapting existing psychotherapy standards to new technologies; it requires AI-specific outcome measures, clearer definitions of therapeutic interventions, standardization of intervention terms, representative research across disorders and populations, and sustained collaboration among mental health professionals, researchers, computer scientists, regulators, and developers. Until that evidence and infrastructure are established, ethical practice requires that AI therapy be approached as an emerging

intervention that should be investigated carefully as soon as possible.

References

- AICBT Ltd. (n.d.). *Overcome social anxiety*. AI-Therapy. <https://www.ai-therapy.com/>
- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). <https://doi.org/10.1176/appi.books.9780890425787>
- Barnes, A. J., Zhang, Y., & Valenzuela, A. (2024). AI and culture: Culturally dependent responses to AI systems. *Current Opinion in Psychology*, 58, 101838. <https://doi.org/10.1016/j.copsyc.2024.101838>
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3), 252–260. <https://doi.org/10.1037/h0085885>
- Cho, Y. M., Rai, S., Ungar, L., Sedoc, J., & Guntuku, S. C. (2023). An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11346–11369. <https://doi.org/10.18653/v1/2023.emnlp-main.698>
- Giacobello, M. L. (2025). Informed consent and bioethical advances in clinical settings. *Frontiers in Psychology*, 16, Article 1654586. <https://doi.org/10.3389/fpsyg.2025.1654586>
- Goldberg, S. B., Baldwin, S. A., Riordan, K. M., Torous, J., Dahl, C. J., Davidson, R. J., & Hirshberg, M. J. (2022). Alliance with an unguided smartphone app: Validation of the Digital Working Alliance Inventory. *Assessment*, 29(6), 1331–1345. <https://doi.org/10.1177/10731911211015310>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hua, Y., Xia, W., Bates, D., Hartstein, G. L., Kim, H. T., Li, M., Nelson, B. W., Stromeyer, C., IV, King, D., Suh, J., Zhou, L., & Torous, J. (2025). Standardizing and scaffolding health care AI-chatbot evaluation: Systematic review. *JMIR AI*, 4, e69006. <https://doi.org/10.2196/69006>
- Jabir, A. I., Martinengo, L., Lin, X., Torous, J., Subramaniam, M., & Tudor Car, L. (2023). Evaluating conversational agents for mental health: Scoping review of outcomes and outcome measurement instruments. *Journal of Medical Internet Research*, 25, e44548. <https://doi.org/10.2196/44548>
- Kang, H., & Shin, W. (2016). Do smartphone power users protect mobile privacy better than nonpower users? Exploring power usage as a factor in mobile privacy protection and disclosure. *Cyberpsychology, Behavior, and Social Networking*, 19(3), 179–185. <https://doi.org/10.1089/cyber.2015.0340>

- Kolding, S., Lundin, R. M., Hansen, L., & Østergaard, S. D. (2025). Use of generative artificial intelligence (AI) in psychiatry and mental health care: A systematic review. *Acta Neuropsychiatrica*, 37, e37. <https://doi.org/10.1017/neu.2024.50>
- Kuhail, M. A., Al-Shamaileh, O., Farooq, S., Shahin, H., Abdelzaher, F., & Thomas, J. (2025). A systematic review on mental health chatbots: Trends, design principles, evaluation methods, and future research agenda. *Human Behavior and Emerging Technologies*, 2025, Article 9942295. <https://doi.org/10.1155/hbe2/9942295>
- McCall, H. C., Richardson, C. G., Helgadottir, F. D., & Chen, F. S. (2018). Evaluating a web-based social anxiety intervention among university students: Randomized controlled trial. *Journal of Medical Internet Research*, 20(3), e91. <https://doi.org/10.2196/jmir.8630>
- Tindle, R. (2021). Improving the global reach of psychological research. *Discover Psychology*, 1, 5. <https://doi.org/10.1007/s44202-021-00004-4>
- Villarreal-Zegarra, D., Reategui-Rivera, C. M., García-Serna, J., Quispe-Callos, G., Lázaro-Cruz, G., Centeno-Terrazas, G., Galvez-Arevalo, R., Escobar-Agreda, S., Dominguez-Rodriguez, A., & Finkelstein, J. (2024). Self-administered interventions based on natural language processing models for reducing depressive and anxiety symptoms: Systematic review and meta-analysis. *JMIR Mental Health*, 11, e59560. <https://doi.org/10.2196/59560>
- Wang, H., Zhao, B., Jiao, Y., Liang, C., & Zheng, G. (2026). The moderating effect of national culture on the relationships between antecedents and attitude and intention for AI adoption: A meta-analysis. *Technological Forecasting and Social Change*, 231, 124803. <https://doi.org/10.1016/j.techfore.2026.124803>
- 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. <https://doi.org/10.1145/365153.365168>
- Yang, H., Chang, F., Muroi, F., Liu, Z., Zhang, W., & Cai, J. (2026). Application of artificial intelligence in schizophrenia rehabilitation management: A systematic scoping review. *Translational Psychiatry*, 16, 180. <https://doi.org/10.1038/s41398-026-03872-3>
- Yin, R., Martinengo, L., Smith, H. E., Subramaniam, M., Griva, K., & Tudor Car, L. (2024). The views and experiences of older adults regarding digital mental health interventions: A systematic review of qualitative studies. *The Lancet Healthy Longevity*, 5(11), 100638. <https://doi.org/10.1016/j.lanhl.2024.08.007>
- Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. *Journal of Medical Internet Research*, 27, e78238. <https://doi.org/10.2196/78238>

Bio:

Jan Warner, PhD, LISWS, is a psychotherapist in private practice in Ohio and is retired from the Louis Stokes Cleveland VA Medical Center. Her area of expertise is the treatment of adults experiencing a range of mood disorders. She earned a Master in Social Service Administration from Case Western Reserve University and a PhD in Clinical Psychology from the California Institute of Human Science. She currently serves on the Research Committee of the Association for Comprehensive Energy Psychology (ACEP).

Contact: janwarnerliswsphd@gmail.com

TERMS OF USE

The International Journal of Healing and Caring is distributed electronically as an open access journal, available at no charge. You may print your downloaded copy of this article or any other article for relaxed reading. We encourage you to share this article with friends and colleagues.

The International Journal of Healing and Caring
31907 E Davis Ranch Rd.
Marana, AZ 85658

Email: ijhcjournal@earthlink.net Website: <http://www.ijhc.org>

Copyright © 2026 IJHC.org, Dr. Jan Warner. All rights reserved. No AI scraping or training of this document is allowed.

DISCLAIMER: <http://ijhc.org/disclaimer>