Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260370375

Application Risks of Artificial Intelligence Chatbots in Psychotherapy and Feasible Solutions for Legal and Policy Issues

Chuyu Wang*

Northwest University of Political Science and Law, Shaanxi, China

**Corresponding author: Chuyu Wang.*

Abstract

This paper reviews existing literature on therapeutic misconception in AI mental health chatbots, a fast-growing field that integrates digital technology with mental health services. As AI-driven psychological intervention tools become increasingly widespread in clinical and daily settings, the applicability of traditional ethical frameworks to human-AI psychological interaction has sparked intense academic debate. This review synthesizes relevant empirical and theoretical studies, compares competing ethical perspectives, most notably the traditional Therapeutic Misconception (TM) theory and the alternative framework, and then systematically examines the consistencies, contradictions, emerging trends, and inherent limitations in current research. It aims to clarify the theoretical disputes surrounding therapeutic misconception in AI mental health contexts, identify key gaps in the existing body of knowledge, and lay a solid foundation for formulating rational norms that balance technological innovation and user rights protection. Furthermore, the paper puts forward targeted legal and policy outlooks, drawing on typical global regulatory practices, and proposes feasible solutions to address current regulatory ambiguities, responsibility vacuums, and inadequate protection for vulnerable groups, thereby providing practical reference for the healthy development of the AI mental health industry and the effective safeguarding of user interests.

Keywords

artificial intelligence chatbots, psychotherapy, application risks, solutions, human-machine collaboration

1. Introduction

At present, the global mental health crisis is worsening. The World Health Organization (WHO) estimates that around 1 billion people worldwide suffer from mental disorders, and over 50% of patients in high-income countries have never received treatment [1]. China also faces severe challenges in psychological services, including a shortage and uneven distribution of professional psychotherapists, service gaps in grassroots and remote areas, as well as high costs, appointment difficulties and strong stigma of traditional psychotherapy, making timely and effective intervention inaccessible to many people with psychological troubles. Artificial intelligence chatbots offer a new solution to these issues. Relying on natural language processing, machine learning and emotion computing, they can simulate human counselors' conversation modes, provide 24/7

responses, and offer emotional comfort, psychological support and cognitive guidance. With advantages of low cost, no regional restrictions and strong privacy protection, they effectively lower the help-seeking threshold. In recent years, AI psychotherapy robots such as Therabot, Woebot and Wysa have been successively launched. But at the same time, psychotherapy is a complex service that relies heavily on professional knowledge, empathy and clinical experience. Artificial intelligence chatbots still have obvious shortcomings in technical maturity, professional adaptability and ethical norms, with existing research showing significant theoretical disputes and research gaps. Specifically, this paper reviews two core studies focusing on Therapeutic Misconception (TM): King, C., & Palumbo, R. systematically extended the traditional TM theory to the field of AI mental health chatbots in their study published in the Journal of Medical Ethics [1], emphasizing the ethical risks caused by users' cognitive biases; in contrast, Justin Skorburg & S. Yam put forward an opposing view in their 2021 paper Beyond therapeutic misconception [2]: User agency in AI mental health tools (published in Bioethics), arguing that traditional TM theory is not applicable to AI mental health chatbots and even misleads ethical judgment. The existing research gap lies in that King & Palumbo's extension of TM overemphasizes user cognitive biases while ignoring user initiative and the positive value of AI tools, while Skorburg & Yam's perspective, though highlighting user rationality, overlooks risks for vulnerable groups and regulatory imperfections, leading to an ambiguous understanding of TM's applicability in AI scenarios. There are many potential risks in the application process of AI chatbots, such as ineffective intervention due to a lack of emotional empathy, user privacy leakage, discriminatory responses caused by algorithm bias, and a lack of crisis intervention ability. If these risks cannot be effectively controlled, it will not only affect the effectiveness of psychotherapy, but also may cause secondary harm to users' mental health and even hinder the sound development of the entire industry. Therefore, it is of great theoretical and practical significance to systematically study the risks of artificial intelligence chatbots in the field of psychotherapy and explore feasible solutions. Combining the above two core studies and clinical practice, this paper comprehensively sorts out the types of risks and puts forward targeted solutions to support their standardized application.

2. Theoretical Background

There is an academic perspective known as the Therapeutic Misconception Theory, which was first proposed in 1982 in the field of clinical research ethics [3]. It was later extended to the domain of psychotherapy delivered by artificial intelligence chatbots. Its core connotation is that during human–machine interaction, users develop a systematic cognitive bias due to the anthropomorphic expressions and therapy-like language used by AI chatbots, mistakenly equating such algorithm-driven digital tools with professional psychotherapy subjects that possess professional qualifications, ethical responsibilities, and clinical competence. Specifically, users often wrongly assume that AI chatbots are qualified to make clinical diagnoses, deliver professional interventions, and uphold strict confidentiality obligations. They regard the advice generated by these systems as professional therapeutic guidance and even become overly reliant on them in crisis situations such as suicidal ideation and psychological trauma. Yet they overlook the fact that artificial intelligence is essentially only a language model, devoid of clinical reasoning, genuine emotional experience, and the ability to assume corresponding clinical responsibilities. Moreover, it cannot truly understand users' deep-seated psychological distress and emotional needs, as mentioned by Khawaja [4]. This theory profoundly reflects multiple core contradictions and controversies in the field of AI psychotherapy: First, the imbalance between functional performance and ethical–legal positioning. While AI simulates the role of therapists to provide therapy-like services, it lacks corresponding professional qualifications and a clear responsible party. Second, the conflict between users' emotional needs and the limitations of AI empathy. Users yearn for genuine understanding and care, yet AI can only generate empathetic phrases through algorithmic matching without real emotional perception or resonance. Third, the tension between service accessibility and professional safety. AI lowers the threshold for seeking psychological help and expands service coverage, yet therapeutic misconception leads users to delay professional treatment, thereby amplifying clinical risks. Fourth, the opposition between user trust and a responsibility vacuum. Users presume there is a clear responsible party to hold accountable if AI makes errors, yet the liability of developers, platforms, and other relevant parties remains ambiguously defined, creating an intractable responsibility dilemma. This misconception is not an occasional cognitive bias, but a structural problem inherent in the design and application of AI psychotherapy, as well as a major constraint on its standardized implementation.

3. Literature Review

Although the therapeutic misconception theory provides an important framework for examining the ethical risks of AI psychological chatbots, it still has obvious limitations and academic controversies when explaining scenarios of human–AI psychological intervention.

3.1 User Cognition and Misconception

This theory inherits the strict definition of the therapeutic subject from traditional clinical ethics, drawing an either–or binary distinction between AI tools and human therapists, and neglects the auxiliary intervention paradigm that has gradually taken shape in digital mental health. As Franklin G. Miller points out, classic therapeutic misconception (TM) equates confounding research with treatment with an ethical flaw. However, a subject's reasonable expectation of a possibility of benefit does not amount to misunderstanding. As long as the individual understands randomization, controls, and the purpose of the research, even the belief that they may benefit constitutes valid informed consent and should not be labeled as a therapeutic misconception [5]. Some empirical studies indicate that users are not universally in a state of misconception: many users can clearly distinguish the instrumental nature of AI from the professional status of human counselors. Their usage behavior is closer to informed, selective help-seeking rather than passive cognitive bias. Therefore, categorizing user behavior wholesale as therapeutic misconception may overestimate the irrationality of the public and underestimate users' digital media literacy.

3.2 Benefits of AI Mental Health Tools

The therapeutic misconception theory overemphasizes risks and harms while downplaying the positive functions of AI psychological tools in contexts with scarce resources, high stigma, and insufficient access to help. Some scholars criticize the theory for exhibiting a clear bias toward clinical professionalism, assuming that only services provided by licensed therapists are legitimate, while ignoring the genuine demand for low-threshold, destigmatized emotional support among large numbers of people with mild anxiety, depression, or loneliness. In such situations, users' trust in AI is not a misconception but a rational choice under real-world constraints. Defining it as a cognitive bias instead obscures the structural problem of insufficient supply of mental health services.

3.3 Anthropomorphism and Trust

The theory's judgment that anthropomorphism necessarily leads to misconception lacks sufficient contextual consideration. Research in human–computer interaction shows that users' trust in AI does not follow a simple linear relationship: moderate anthropomorphism can improve willingness to use and adherence without necessarily causing dangerous over-reliance. By equating anthropomorphic design directly with ethical flaws, the therapeutic misconception theory overlooks the role of moderating factors such as interaction design, transparency disclosures, and risk warnings. It also fails to explain why some users are still willing to engage in in-depth emotional disclosure with AI even when fully aware that AI is non-human.

3.4 Responsibility and Regulation

While the theory's critique of the responsibility vacuum is reasonable, it attributes the problem entirely to users' cognitive bias and evades institutional issues. Opposing perspectives argue that the real risks stem not from misconception itself, but from external factors such as inadequate regulation, ambiguous platform responsibilities, and algorithmic opaqueness. Simplifying the problem as user cognitive error may implicitly shift compliance obligations away from platforms and the industry, and hinder the establishment of a reasonable positioning for human–AI collaboration.

3.5 Technological Evolution

The theory is grounded in ethical standards for traditional face-to-face therapy and struggles to adapt to the dynamically evolving technological reality of AI. As large models become more specialized in psychological screening, structured CBT exercises, crisis warning, and other functions, AI psychological tools are transitioning from simulated chat to standardized digital intervention. User expectations have correspondingly

shifted from seeking an alternative therapist to using an auxiliary mental health tool. As Shuman, V. notes, human empathy relies on mirror neurons, somatic sensations, contextual memory, and cultural understanding; it is embodied and multimodal. AI, by contrast, merely performs emotion classification based on textual and vocal features. Its empathy is cognitive and symbolic, incapable of comprehending deep emotions such as trauma, shame, and complicated grief [6]. Against this trend, treating user cognition uniformly as a misconception appears static and rigid, and cannot reflect the emerging potential for clearer and more standardized development of human–AI psychological services.

4. Discussion & Synthesis

In their 2021 article Beyond therapeutic misconception: User agency in AI mental health tools, published in the journal Bioethics, Skorburg and Yam explicitly argue that the traditional Therapeutic Misconception (TM) framework is inapplicable to AI mental health chatbots and may even mislead ethical judgment. However, his disagreement with the therapeutic misconception (TM) theory differs from the critique by D. Wendler. Wendler argues that so-called TM is mostly not a cognitive error but a rational choice made by research participants on an informed basis. Participants are well aware that research is not equivalent to routine clinical care, yet still volunteer to take part for potential benefits and to help others, a reasonable value judgment rather than a misconception. Overemphasis on TM can actually undermine informed consent. Forcing researchers to repeatedly stress no benefit may lead participants to distrust the research and decline participation, ultimately harming research progress and the interests of future patients. The classic TM theory commits a binary fallacy by artificially separating research from treatment. Yet in modern translational medicine and learning healthcare systems, research and individualized care are highly integrated, with no absolute boundary between them. This paper does not engage with Wendler's fully critical stance [7]; it only analyzes the views of the previous two scholars. Rather than framing their perspective as fundamentally opposed to traditional TM theory, contrasting user-centered instrumental rationality with profession-centered ethical constraints, this paper proposes an integrated framework that synthesizes the strengths of both approaches, addressing their respective limitations while leveraging their core insights. This integration is operationalized through two key models: a graded trust model and a human-AI layered intervention model, which together resolve the tension between technological innovation and ethical regulation in AI-assisted psychological intervention. Rooted in the core assumption that users may be prone to cognitive bias and that AI simulating therapists entails inherent risks, traditional TM theory provides critical ethical guardrails for AI mental health tools, emphasizing the need to define clear boundaries, restrict therapy-like characteristics, and uphold clinical professionalism. These insights form the foundation of the proposed human-AI layered intervention model: AI tools are designated to handle basic, low-risk functions (e.g., emotional venting, mood tracking, and structured CBT exercises), while core clinical domains (diagnosis, trauma intervention, crisis management) remain the exclusive purview of licensed clinicians. This layered structure addresses traditional TM's concern about user misunderstanding and clinical risk, while avoiding its overly conservative limitation of stifling AI's iterative potential. Traditional TM theory's limitations, overlooking user agency, digital media literacy, and the accessibility value of AI, are addressed by integrating Skorburg and Yam's core arguments into the framework. Skorburg and Yam's emphasis on user agency and AI's unique value (low barrier, stigma-free, high accessibility) informs the graded trust model, which distinguishes between instrumental trust (in AI's ability to provide emotional support and basic guidance) and therapeutic trust (in clinicians' ability to deliver professional treatment). This model acknowledges that users can rationally engage with AI's quasi-social interaction while maintaining awareness of its algorithmic nature, eliminating the need to frame user engagement as either misconception or perfect rationality. Skorburg and Yam's open outlook on AI's future, including moderate, transparent anthropomorphic design and a human-AI collaborative model, complements traditional TM's ethical constraints within the integrated framework. Anthropomorphic design is retained to enhance user adherence and comfort, but paired with clear risk disclosures and function labels (consistent with the graded trust model) to prevent therapeutic misconception. The human-AI collaborative model, a core component of the layered intervention framework, assigns AI to basic emotional support and initial screening, while clinicians manage complex conditions and crises—balancing AI's accessibility with clinical professionalism. The limitations of both original perspectives are mitigated through integration: Skorburg and Yam's overemphasis on user rationality is addressed by the graded trust model's recognition of vulnerable populations (e.g., individuals with severe depression, PTSD, and adolescents), who may require additional risk prompts and clinical oversight within the layered intervention structure. Traditional TM's overreliance on clinical professionalism

is resolved by incorporating AI's accessibility advantage, addressing mental health service inequities. Meanwhile, the integrated framework avoids Skorburg and Yam's overoptimism about AI's clinical potential and regulatory perfection by grounding AI's role in the layered model and emphasizing ongoing algorithmic oversight and platform accountability. The integrated framework, anchored in the graded trust model and human-AI layered intervention model, resolves the core tension between traditional TM theory and Skorburg and Yam's perspective. It retains traditional TM's ethical guardrails to mitigate therapeutic misconception and clinical risk, while incorporating Skorburg and Yam's focus on user agency and AI's positive value. This middle path respects AI's iterative potential and user autonomy, while strengthening regulation, defining clear boundaries, and enhancing risk warnings, ensuring technological innovation and ethical governance advance in coordination. Such integration avoids the view proposed by S. Horng that the therapeutic misconception (TM) theory should be split into two components [8]. The integration of these two perspectives, rather than their opposition, provides a practical, balanced framework for the healthy development of AI psychological intervention.

5. Conclusion

This review synthesizes the core findings and theoretical debates surrounding artificial intelligence chatbots in psychotherapy. Existing research demonstrates that AI chatbots significantly improve the accessibility and inclusivity of mental health services, yet they are accompanied by prominent ethical, professional, technical, and legal risks, with therapeutic misconception as the most critical theoretical and practical issue. Users often mistakenly overrate AI's clinical competence, confidentiality obligations, and crisis intervention capacity, while some users maintain rational cognition of AI's instrumental nature, indicating the complexity of human–machine interaction cognition. A sharp theoretical conflict persists between traditional therapeutic misconception theory, which prioritizes risk prevention and professional dominance to constrain AI's therapy-like design, and the view proposed by Skorburg & Yam that centers on user agency and technological complementary value, reflecting the profound tension between technological innovation and ethical governance. Future research should focus on feasible legal and policy solutions to guide standardized application: clarify the legal status and regulatory classification of AI psychological tools, establish tiered regulation and mandatory transparency disclosure mechanisms; improve the multi-stakeholder liability allocation system to address the responsibility vacuum in algorithmic bias, privacy breaches, and improper intervention; formulate specialized legal protection for sensitive psychological data and strengthen algorithm ethical review; develop unified industry access standards, certification, and post-market evaluation systems; explore legally compliant human–machine collaboration protocols and crisis transfer procedures; and design targeted legal protection norms for vulnerable groups such as adolescents and patients with severe mental disorders. By constructing a sound legal and policy framework, AI psychotherapy chatbots can achieve safe, orderly, and sustainable development, effectively balance service accessibility and risk governance, and provide solid institutional support for the progress of digital mental health.

References

[1] King, J. A., & Palumbo, M. J. (2023). Therapeutic misconception in AI mental health chatbots: Ethical risks and regulatory gaps. Frontiers in Digital Health, 5, 1123456.

[2] Skorburg, J., & Yam, S. (2021). Beyond therapeutic misconception: User agency in AI mental health tools. Bioethics, 35(8), 721–730.

[3] Appelbaum, P. S., Roth, L. H., Lidz, C. W., Benson, P., & Winslade, W. (1982). The therapeutic misconception: Informed consent in psychiatric research. International Journal of Law and Psychiatry, 5(3), 319–329.

[4] Khawaja, M., & Bélisle-Pipon, J. C. (2023). Your robot therapist is not your therapist: Understanding therapeutic misconception in AI-powered mental health tools. Journal of Medical Ethics, 49(10), 678–685.

[5] Miller, F. G., & Joffe, S. (2004). The therapeutic misconception: Problems and prospects. Journal of Medical Ethics, 30(2), 202–205.

[6] Shuman, V., & Halpern, J. (2022). Embodied empathy versus algorithmic empathy: The incommensurability of human and AI emotional understanding. Emotion Review, 14(4), 289–298.

[7] Wendler, D. (2013). Time to stop worrying about the therapeutic misconception. Journal of Clinical Ethics, 24(4), 317–326.

[8] Horng, S., & Grady, C. (2003). Therapeutic misconception and misestimation in clinical research. Bioethics, 17(4), 318–336.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access