Gac Méd Caracas 2025;133(4):1250-1256
DOI: 10.47307/GMC.2025.133.4.26

Artificial Intelligence in Psychotherapy: Reflections on Opportunities, Challenges, and Ethical Implications

Inteligencia artificial en psicoterapia: reflexiones sobre oportunidades, desafíos e implicaciones éticas

Lorena Cudris-Torres[1], Maira A. López-Castellar[2], Jenny Danna-Buitrago[3]

SUMMARY

Artificial intelligence (AI) has rapidly become a significant influence in healthcare, and mental health services are no exception. In psychotherapy, technologies ranging from conversational agents (chatbots) and social assistance robots to immersive virtual reality platforms for clinical support, symptom management, and therapeutic augmentation are being developed and evaluated. This reflective article examines recent advances in AI applied to psychotherapy. It discusses promising clinical prospects (improved access, scalability, and novel therapeutic modalities), alongside key limitations such as uncertain long-term effectiveness, risks to the therapeutic alliance, algorithmic bias, data governance, and regulatory gaps. Concrete examples (Woebot, Wysa, Tess, Paro, NAO, bubble-based therapies, and avatars) are used to illustrate the benefits and risks and to inform a set of practical recommendations for clinicians, developers, and policymakers. The article concludes that the responsible integration of AI into psychotherapy requires robust evidence (randomized trials and long-term follow-up), clear ethical and legal frameworks, clinical oversight, and active efforts to preserve human-centered therapeutic values. We propose research and policy priorities that seek to maximize benefits and minimize harm.

Keywords: *Artificial intelligence, psychotherapy, chatbots, robots, virtual reality, ethics.*

RESUMEN

La inteligencia artificial (IA) se ha convertido rápidamente en una fuerza influyente en la atención médica, y los servicios de salud mental no son la excepción. En psicoterapia, se están desarrollando y evaluando tecnologías que abarcan desde agentes conversacionales (chatbots) y robots de asistencia social hasta plataformas de realidad virtual inmersiva para el apoyo clínico, el manejo de síntomas y la potenciación terapéutica. Este artículo reflexivo examina los avances recientes en la IA aplicada a la psicoterapia. Analiza las prometedoras perspectivas

DOI: https://doi.org/10.47307/GMC.2025.133.4.26

ORCID: 0000-0002-3120-4757[1*]
ORCID: 0000-0002-9761-840X[2]
ORCID: 0000-0003-0241-9481[3]

[1]Universidad de la Costa, Barranquilla, Colombia. E-mail: lcudris3@cuc.edu.co

Recibido: 8 de septiembre 2025
Aceptado: 30 septiembre 2025

[2]Fundación Universitaria del Área Andina, Valledupar, Colombia. E-mail: malopez57@areandina.edu.co
[3]Fundación Universitaria del Área Andina, Bogotá, Colombia. E-mail: jdanna@areandina.edu.co

*Corresponding author: Lorena Cudris-Torres, Senior Lecturer 3, Department of Social Sciences, Universidad de la Costa, lcudris3@cuc.edu.co

clínicas (mejor acceso, escalabilidad y nuevas modalidades terapéuticas), junto con limitaciones clave como la incertidumbre sobre la efectividad a largo plazo, los riesgos para la alianza terapéutica, el sesgo algorítmico, la gobernanza de datos y las lagunas regulatorias. Se utilizan ejemplos concretos (Woebot, Wysa, Tess, Paro, NAO, terapias basadas en burbujas y avatares) para ilustrar los beneficios y riesgos, y para fundamentar un conjunto de recomendaciones prácticas para profesionales clínicos, desarrolladores y legisladores. El artículo concluye que la integración responsable de la IA en la psicoterapia requiere evidencia sólida (ensayos aleatorizados y seguimiento a largo plazo), marcos éticos y legales claros, supervisión clínica y esfuerzos activos para preservar los valores terapéuticos centrados en el ser humano. Proponemos prioridades de investigación y políticas para maximizar los beneficios y minimizar los daños.

Palabras clave: *Inteligencia artificial, psicoterapia, chatbots, robots, realidad virtual, ética.*

INTRODUCTION

In recent years, the field of psychotherapy has witnessed the emergence of a broad set of artificial intelligence (AI) applications designed to support or augment therapeutic work. These applications encompass conversational agents that offer psychoeducation and guided exercises, socially assistive robots that provide companionship and stimulation, and immersive virtual reality systems that facilitate graded exposure and avatar-mediated interaction. A recent monograph offers a comprehensive overview of these developments and their empirical foundation, providing a useful point of departure for critical reflection (1).

The COVID-19 pandemic accelerated the adoption of remote and digital mental health tools, renewing interest in scalable and accessible interventions. Global policy documents and guidance emphasise the potential of digital health to reduce unmet need while signaling the necessity of governance, standards, and evaluation frameworks to safeguard users and health systems (2,3). In mental health, where relational factors and trust are central, AI presents both opportunity and substantial ethical complexity. This paper presents a reflective analysis that maps current technologies to core psychotherapeutic processes, evaluates the existing empirical evidence, and offers practical recommendations for clinicians, researchers, and policymakers.

In this regard, the documents supporting the reflections were obtained through a narrative literature review, with an emphasis on indexed articles, systematic reviews, and reports from international organizations. Priority was given to sources published between 2017 and 2024, retrieved from academic databases such as PubMed, Scopus, and Web of Science, as well as official documents from the World Health Organization (WHO) and regulatory bodies.

The aims of this reflection are threefold: (1) to synthesise recent evidence and representative case-studies of AI in psychotherapy; (2) to critically assess benefits, limitations, and ethical challenges; and (3) to propose pragmatic priorities for research, deployment, and regulation that preserve the therapeutic values that underpin effective psychotherapy.

AI and psychotherapy: opportunities and challenges

Artificial intelligence in psychotherapy can be conceptualized along a continuum that ranges from decision-support tools that assist clinicians to fully automated therapeutic agents that interact directly with users. Tools that augment clinician practice, such as algorithms for risk stratification, automated outcome measurement, or session summarization, can increase efficiency and support personalized planning. Fully automated interventions (chatbots, virtual therapists) offer novel routes to reach people who would otherwise not access services due to cost, stigma, or limited provider supply (1,4).

Key opportunities include improved accessibility (24/7 availability and lower marginal cost), consistent delivery of structured interventions (useful for brief Cognitive Behavioral Therapy (CBT)-style exercises), and the capacity to collect and analyze large amounts of user data to detect early warning signs and personalized interventions (5,6,8). Some technologies (for example, virtual avatars or socially assistive robots) enable therapeutic practices that are difficult or impossible in

conventional talk therapy, graded, fully controlled exposures, and embodied interaction that can be tailored to sensory and cognitive profiles.

At the same time, challenges are substantial. Evidence for many AI interventions remains limited: trials are often small, short in duration, and focused on symptom reduction rather than functional or relational outcomes (5,7,8). The therapeutic alliance, an established predictor of psychotherapy outcomes, may be altered in AI-mediated care. Some studies report surprising levels of alliance with chatbots, while others highlight superficiality and the risk of misleading users regarding their capabilities (15,13). Algorithmic bias, opaque model behavior, data privacy risks, and unclear lines of clinical responsibility pose practical and ethical barriers to the uptake of these models (3,13). In addition, the predominance of CBT-like approaches in digital products raises questions about the pluralism of psychotherapeutic models embedded in AI and the cultural relevance of training data and design teams.

Chatbots: promise and limitations

Conversational agents (chatbots) have been among the most visible AI interventions in mental health. Notable examples include Woebot, Wysa and Tess, each designed with different clinical emphases and modes of deployment (5-7). Woebot was created as an automated agent informed by cognitive behavioral therapy and was assessed in a randomized trial among young adults; the study found it to be feasible, acceptable, and associated with short-term reductions in depression and anxiety symptoms (5). Wysa, evaluated at scale using real-world usage data and mixed-methods designs, demonstrates utility for self-management and shows promising user engagement patterns. Studies of therapeutic alliance with Wysa indicate that users can personify and form a working alliance with a chatbot, even when aware of its non-human nature (6,15). Tess has been trialled in targeted clinical contexts, and feasibility studies show high satisfaction and potential symptom effects in specific samples (7).

Systematic reviews and meta-analyses of mental health chatbots highlight a mixed but cautiously optimistic picture: chatbots can reduce symptoms in short-term studies, improve engagement for some users, and serve as adjunctive tools; however, heterogeneity of trials, lack of long-term follow-up, and limited replication undermine firm conclusions (8).

Limitations are important. First, chatbots often rely on scripted or semi-structured CBT techniques; for users with complex or comorbid presentations, scripted responses may be insufficient and could even delay access to more appropriate care. Second, natural language understanding remains imperfect: chatbots struggle with sarcasm, metaphors and highly idiosyncratic narratives, and this can lead to misinterpretation or inappropriate suggestions (5,6,8). Third, safety and escalation protocols vary widely across products: not all systems reliably detect suicidal ideation or acute risk, and there is an inconsistent connection to human crisis services. Finally, transparency and informed consent are often inadequately addressed in consumer-facing apps.

Despite these caveats, chatbots occupy an important niche: providing low-intensity, scalable support for mild to moderate problems, offering guided self-help between sessions, and serving as triage functions that can direct high-risk individuals to human care. For clinicians and services considering chatbots, priorities include: (a) selecting products with transparent safety protocols and published evaluations; (b) integrating chatbot data streams with clinician oversight rather than offloading responsibility; and (c) monitoring for differential engagement and outcomes across demographic groups to guard against inequitable effects.

Robotherapy: ethical and clinical reflections

Socially assistive robots (often described as 'robotherapy') constitute a distinct strand of AI interventions. Paro a robotic therapeutic seal and NAO a small humanoid robot used in autism interventions are among the most studied platforms (10,11,16,17). Paro is designed to provide sensory stimulation and companionship in care settings. Scoping reviews and controlled studies report reductions in agitation, improvements in mood, and increased social engagement

among older adults with dementia, along with physiological correlations, such as reduced heart rate in some samples (10,16,17).

NAO and similar humanoid platforms have been used in interventions for children with autism spectrum disorders (ASD). Studies suggest that predictable, programmed interactions with NAO can scaffold joint attention, turn-taking, and simple social skills in some children with ASD, possibly by reducing the social unpredictability of human interaction (11).

Ethical and practical issues are prominent in robotherapy. First, the cost and maintenance of physical robots limit scalability in many settings. Second, staff training and acceptability are variable; caregivers may lack confidence to deploy robots effectively, and institutions may lack budgets. Third, there are moral questions about simulating affective relations: when a vulnerable person forms an attachment to a robot, is this ethically acceptable if the robot cannot reciprocate? Advocates argue that therapeutic benefits can reduce loneliness, increase engagement, and be ethically justifiable if they are real, transparent, and supervised by staff. Opponents caution against replacing human contact and call for a thorough evaluation of the longer-term effects and potential unintended consequences.

Virtual reality and immersive interventions

Immersive virtual reality (VR) combined with AI opens further therapeutic possibilities. VR has long-standing evidence for specific applications (for example, exposure therapy for phobias) and contemporary systems pair VR scenarios with AI-driven personalization and conversational guidance (9). Bubble, a VR+AI app developed for women with cancer experiencing hot flashes and distress, reported reductions in hot-flash frequency and improvements in psychological well-being in a pilot study, highlighting the potential of tailored immersive environments supported by AI-guided coaching (9).

Avatar-mediated therapies for persistent auditory hallucinations are particularly interesting: here, clinicians and technologists co-create an avatar that represents the voice the patient hears; the therapeutic process then externalizes the persecutory voice, allowing the patient to practice response strategies in a controlled, graded context. Controlled and feasibility studies report improvements in voice-related distress and reduced frequency in some cohorts, although large-scale trials and replication are needed (12).

VR+AI interventions demand specialized infrastructure and clinician training, and they carry risks such as cybersickness and emotional destabilization if not appropriately paced and supervised. These interventions may nevertheless provide powerful clinical tools, especially when integrated with human-led therapy and clear safety protocols.

Cross-cutting ethical dilemmas

Across modalities, several ethical themes recur. Data governance and privacy are paramount: mental health data are highly sensitive, and AI systems typically require large datasets for training and ongoing personalization. Strong safeguards, minimal datasets, secure storage, and clear user consent are non-negotiable (3).

Transparency and explainability carry both practical and ethical weight. Many contemporary AI systems (large language models, deep neural networks) are 'black boxes' in their internal representations. For clinicians and patients to trust recommendations or therapeutic prompts, developers must provide clear descriptions of system capabilities and limitations, as well as robust safety-testing results (3,13).

Equity and bias demand active attention. Training data reflect cultural norms and health systems; models trained largely on English-speaking, high-income populations may perform poorly in other contexts, risking misdiagnosis or poor engagement in underrepresented groups. Regulators and developers should require evidence of cross-cultural validity and fund translation/localization efforts (2,3,14).

Therapeutic responsibility and liability are thorny: when an AI agent provides advice or fails to detect risk, who is accountable—the developer, the clinician who recommended the app, or the service that integrated it? Clear clinical governance frameworks, mandatory escalation pathways for risk, and role definitions

are essential to avoid gaps in care and ambiguous liability (3,13).

Finally, preserving human-centred therapeutic values, empathy, containment, and meaning-making must be a guiding principle. AI can augment but should not trivialise relational work. When automation is used, clinicians should be mindful of how AI influences patient narratives and should maintain space for human reflection and care.

DISCUSSION

The evidence landscape for AI in psychotherapy is maturing but remains uneven. Several high-quality small trials and numerous feasibility and real-world evaluations show promising short-term effects for symptom reduction and acceptability across a range of tools (chatbots, VR, robotic companions) (5-9,10,12). However, heterogeneity in study designs, variable outcome measures, and limited long-term follow-up caution against over-optimistic claims. It is also important to note that the research designs reviewed present significant limitations: pilot and feasibility studies predominate, often with small samples and short follow-up periods; heterogeneity in evaluation criteria is common, and active comparators are frequently absent. These limitations restrict the generalizability of the findings and highlight the need for larger and more representative clinical trials that include functional, relational, and cultural outcome measures. The field needs larger pragmatic trials, replication across diverse populations, and standardized outcome frameworks that capture symptom change, functional outcomes, and relational metrics such as alliance.

For clinicians, thoughtful adoption is required. Recommended practical steps are as follows: (a) evaluate the evidence base and safety features of any tool prior to making recommendations; (b) utilize AI tools to supplement, rather than replace, human-led care whenever feasible; (c) ensure that informed consent protocols clearly communicate limitations and data usage; and (d) record AI-generated data in clinical documentation in accordance with standard data protection procedures (3,18).

For developers and researchers, co-designing with clinicians and patients is crucial to ensure usability, cultural relevance, and safety. Trial designs should incorporate active comparators, extended follow-up periods, and mixed-method evaluations that examine both subjective experiences and relational processes. Regulators should require transparent reporting, post-market surveillance, and clear labelling of intended use and limitations.

Policymakers should prioritize equitable access, infrastructure support for low-resource settings, and funding for independent comparative effectiveness research. International guidance (for example, the WHO digital health strategy and WHO AI ethics guidance) offers a foundation for harmonized standards, but national regulatory clarity remains crucial (2,3).

CONCLUSIONS

For clinicians, the practical implications lie in carefully integrating AI tools as adjuncts rather than substitutes for therapeutic work. Practitioners should be equipped to critically evaluate digital interventions, assess their evidence base, and communicate transparently with patients about the potential benefits and limitations of AI-mediated care. Central to this process is safeguarding the therapeutic alliance and ensuring clear boundaries of professional responsibility.

For researchers, the challenge is to design robust studies that move beyond feasibility trials. Future investigations should incorporate long-term follow-up, comparative effectiveness analyses against standard care, and culturally sensitive approaches that reflect diverse populations. Mixed-methods designs are especially valuable in capturing not only symptom change but also relational, ethical, and experiential aspects of AI in psychotherapy.

For policymakers and regulators, the priority is to create frameworks that foster innovation while protecting patient rights and interests. This includes establishing national standards for digital mental health, mandating transparency in algorithmic processes, and ensuring equitable access to care regardless of socioeconomic

status or geographic location. Sustainable implementation will depend on partnerships between governments, academic institutions, clinicians, and technology developers to ensure safe, ethical, and culturally responsive deployment of AI in psychotherapy.

Conflict of interest

The author declares no conflicts of interest related to the preparation of this reflective article.

REFERENCES

1. González Larrondo A. Aplicación de inteligencia artificial en procesos psicoterapéuticos. Trabajo Final de Grado. Facultad de Psicología, Universidad de la República; 2023.
2. World Health Organization. Global strategy on digital health 2020–2025. Geneva: World Health Organization; 2020.
3. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: World Health Organization; 2021.
4. Fiske A, Henningsen P, Buyx A. The ethical challenges of social robots in care settings. J Med Internet Res. 2019;21(5):e13216.
5. 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 randomised controlled trial. JMIR Ment Health. 2017;4(2):e19.
6. 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 Uhealth. 2018;6(11):e12106.
7. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomised controlled trial. JMIR Ment Health. 2018;5(4):e9782.
8. Montenegro JLZ, da Silva VS, Loureiro LMF, Faustino AM. A systematic review of chatbots in mental health: Benefits and challenges. Psychiatry Res. 2020;292:113365.
9. Horesh D, Kohavi S, Shilony-Nalaboff L, Rudich N, Greenman D, Feuerstein JS, Abbasi MR. Virtual reality combined with artificial intelligence (VR-AI) reduces hot flashes and improves psychological well-being in women with breast and ovarian cancer: A pilot study. Healthcare (Basel). 2022;10(11):2261.
10. Hung L, Liu C, Woldum E, Au-Yeung A, Berndt A, Wallsworth C, et al. The benefits of and barriers to using a social robot PARO in care settings: A scoping review. BMC Geriatr. 2019;19:356.
11. Ismail LI, Shamsudin S, Yussof H, Hanapiah FA, Zahari NI. Robot-based intervention program for autistic children with humanoid robot NAO: Initial response in stereotyped behaviour. Procedia Eng. 2012;41:1441-1447.
12. Du Sert M, Potvin S, Dumais A. Avatar-based and virtual reality therapies for auditory hallucinations: Feasibility and pilot data. (See: Craig et al.; Du Sert et al. in literature). 2018.
13. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent?. Am J Bioeth. 2022;22(9):1-10.
14. Terra M, Baklola M, Ali S, El-Bastawisy K. Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: A narrative review. Egypt J Neurol Psychiatry Neurosurg. 2023;59:1-10.
15. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study. Front Digit Health. 2022;4:847991.
16. Manso-Herrero J, Lacomba-Trejo L. Use of PARO robot in geriatric care: review and clinical perspectives. 2020.
17. Wada K, Ikeda Y, Inoue K, Uehara R. Development and preliminary evaluation of robot therapy using the therapeutic seal robot Paro. In: Proceedings of the 19th International Symposium on Robot and Human Interactive Communication; 2010.p.470-475.
18. Torous J, Kiang MV, Lorme J, Onnela JP, Newby JM. Connected and open, but not too much: a survey of attitudes towards mental health and digital technology. Digital Health. 2018;4:2055207618778572.
19. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 3rd edition. Pearson; 2010.
20. Luxton DD. Artificial intelligence in psychological practice: Current and future applications and implications. Prof Psychol Res Pract. 2014;45(5):332-339.
21. Hill J, Ford WR, Farreras IG. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. Comput Hum Behav. 2015;49:245-250.
22. Romero M, Casadevante C, Montoro H. Cómo construir un psicólogo-chatbot. Papel Psicól. 2020;41(1):27-34.

23. Shibata T, Coughlin JF. Trends of robot therapy with the neurological therapeutic seal robot, PARO. J Robot Mechatron. 2014;26(4):418-425.

24. Hardy T. IA (inteligencia artificial). Polis Rev Latinoam. 2001;(2):18.

25. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024. Disponible en: https://apps.who.int/iris/handle/10665/376104