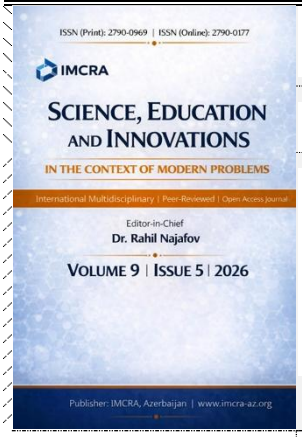

Science, Education and Innovations in the Context of Modern Problems
Issue 5, Vol. 9, 2026

RESEARCH ARTICLE

Reframing Ethical Governance of Artificial Intelligence in Mental Health Care: Toward a Human-Centered, Explainable, and Clinically Responsible Psychotherapeutic Paradigm

Senoussaoui, Abderrahmen
PhD, Clinical Psychology
University of Mohamed Ben Ahmed Oran 2
Algeria
E-mail: senoussaoui13@gmail.com

Rahmani, Djamel
PhD, School Psychology
Mouloud Mammeri University of Tizi Ouzou
Algeria
E-mail: rahmanidjamel13@gmail.com

Keywords
Artificial Intelligence in Mental Health; Ethical Governance; Psychotherapy and Digital Health; Algorithmic Bias and Fairness; Explainable AI (XAI); Patient Autonomy and Data Privacy; Human–AI Interaction; Digital Phenotyping; Clinical Decision-Making; Ethics of Care Framework

Abstract
The rapid integration of Artificial Intelligence (AI) into Mental Health care represents a transformative shift in the diagnosis, monitoring, and treatment of psychological disorders. While AI-driven tools—including digital phenotyping systems, natural language processing models, and conversational agents—offer significant potential to enhance diagnostic precision, accessibility, and personalized interventions, they simultaneously introduce complex ethical, clinical, and epistemological challenges. This study provides a comprehensive and critical synthesis of contemporary literature (2019–2025) to examine the ethical implications of AI deployment in psychotherapeutic contexts. Moving beyond descriptive review, the paper develops a conceptual ethical governance framework grounded in the principles of autonomy, beneficence, non-maleficence, and justice, as well as the emerging "ethics of care" paradigm. It identifies key risk domains, including algorithmic bias, opacity of decision-making processes, data privacy vulnerabilities, erosion of therapeutic alliance, and the reconfiguration of professional accountability. Furthermore, the study proposes a human-centered, explainable, and clinically supervised AI integration model, emphasizing transparency, continuous ethical auditing, stakeholder inclusion, and hybrid human–AI decision-making structures. The findings highlight that while AI can significantly improve efficiency and expand access to mental healthcare—particularly in resource-constrained settings—its unregulated or poorly designed application may exacerbate inequalities, compromise patient autonomy, and undermine trust in clinical practice. The paper contributes to the interdisciplinary discourse by offering a structured ethical framework that bridges technological innovation with clinical responsibility and humanistic values. It concludes that sustainable and ethically aligned AI integration in mental health requires robust regulatory architectures, cross-disciplinary collaboration, and the preservation of the fundamentally human dimensions of psychotherapy.

Citation
Senoussaoui, A., Rahmani, D. (2026). Reframing Ethical Governance of Artificial Intelligence in Mental Health Care: Toward a Human-Centered, Explainable, and Clinically Responsible Psychotherapeutic Paradigm. *Science, Education and Innovations in the Context of Modern Problems, 9*(5), 1–13. https://doi.org/10.56334/sei/9.5.22

Received: September 11, 2025 | **Accepted:** February 22, 2026 | **Published Online:** April 12, 2026

Introduction

Mental disorders represent one of the most critical and rapidly escalating global public health challenges, exerting substantial pressure on healthcare systems, psychiatric institutions, and socio-economic structures worldwide. The growing prevalence of conditions such as depression, anxiety disorders, and behavioral pathologies has been accompanied by a persistent and systemic shortage of qualified mental health professionals, including psychiatrists, clinical psychologists, and psychosocial practitioners. This structural imbalance has resulted in prolonged waiting times for diagnosis and treatment, which not only diminish the overall quality and timeliness of care but also contribute to the deterioration of patients' clinical conditions and long-term health outcomes. Empirical evidence indicates that delayed access to psychological services is strongly associated with worsening symptom severity, particularly in cases of major depressive disorders, thereby reinforcing the urgency of systemic innovation in mental healthcare delivery.

Within this context, the emergence and rapid advancement of Artificial Intelligence (AI) has introduced a transformative paradigm shift in the organization, accessibility, and epistemological foundations of Mental Health services. Positioned at the intersection of computational science and clinical practice, AI-driven technologies—such as machine learning algorithms, natural language processing systems, and digital phenotyping tools—offer unprecedented opportunities to enhance diagnostic precision, enable real-time patient monitoring, and deliver personalized therapeutic interventions at scale. These developments are particularly significant in the context of the Fourth Industrial Revolution, where data-driven decision-making and automation increasingly shape healthcare infrastructures. However, despite these promising advancements, the integration of AI into psychotherapeutic and counseling practices introduces a complex and multidimensional set of ethical, clinical, and epistemological challenges. Core concerns include algorithmic bias and discrimination, risks to patient privacy and data security, lack of transparency in "black-box" decision-making systems, and the potential erosion of the therapeutic alliance—a foundational element of effective psychological treatment. Moreover, the increasing reliance on automated systems raises fundamental questions regarding professional accountability, patient autonomy, and the preservation of inherently human dimensions of care, such as empathy, trust, and interpersonal communication. These issues highlight the tension between technological efficiency and ethical responsibility, necessitating a critical reassessment of existing frameworks governing mental healthcare. While the current body of literature has extensively documented both the opportunities and risks associated with AI in mental health, there remains a notable gap in the development of integrative, human-centered ethical governance models that systematically align technological innovation with clinical responsibility and patient-centered values. In response to this gap, the present study advances a conceptual and analytical framework that reinterprets the ethical governance of AI in psychotherapeutic contexts through the lens of both classical bioethical principles and contemporary relational approaches, including the ethics of care.

Accordingly, this paper aims to (i) critically examine the principal ethical challenges arising from the deployment of AI in mental health diagnosis and therapy, (ii) synthesize existing theoretical and empirical insights into a coherent analytical structure, and (iii) propose a human-centered, transparent, and clinically supervised model for the responsible integration of AI technologies in psychotherapy. By doing so, the study contributes to the evolving interdisciplinary discourse at the intersection of technology, ethics, and mental healthcare, emphasizing that the sustainable future of AI in this domain depends not only on technical sophistication but also on the preservation of fundamental human values and ethical integrity.

Literature Review

The rapid expansion of Artificial Intelligence in Mental Health care has generated a substantial and interdisciplinary body of literature spanning clinical psychology, data science, ethics, and health policy. Existing research can be broadly categorized into three interrelated domains: (i) technological applications of AI in mental healthcare, (ii) ethical and socio-technical challenges, and (iii) emerging governance frameworks.

1. AI Applications in Mental Health Care

A growing number of studies highlight the transformative potential of AI-driven technologies in improving diagnostic accuracy, treatment personalization, and accessibility of care. Machine learning models, natural language processing (NLP), and digital phenotyping have been widely explored as tools for early detection and continuous monitoring of mental disorders (D'Alfonso, 2020; Lee et al., 2022). In particular, conversational agents and AI-supported cognitive behavioral therapy (CBT) systems have demonstrated promising outcomes in reducing symptoms of anxiety and depression while increasing access to psychological support in underserved populations (Beg et al., 2024).

Systematic reviews further confirm that AI applications are increasingly being deployed across diagnostic, predictive, and therapeutic domains, with significant advancements in risk prediction, behavioral pattern recognition, and personalized intervention design (Cruz-Gonzalez et al., 2025). However, despite these advancements, the clinical integration of AI remains uneven and often constrained by data limitations, lack of standardization, and challenges related to interpretability.

2. Ethical and Clinical Challenges

Parallel to technological progress, a substantial body of literature has critically examined the ethical implications of AI in mental healthcare. Core concerns consistently identified across studies include algorithmic bias, data privacy risks, lack of transparency in decision-making, and the erosion of the therapeutic relationship (Fiske et al., 2019; Farmer et al., 2024). These challenges are particularly pronounced in mental health contexts, where data are highly sensitive and clinical decisions rely heavily on subjective and relational dimensions.

Recent empirical and review-based studies emphasize that AI systems may inadvertently reproduce or amplify existing social inequalities due to biased training datasets, leading to disparities in diagnosis and treatment outcomes (Saeidnia et al., 2024). Additionally, the "black-box" nature of many machine learning models raises concerns regarding explainability and accountability, especially in high-stakes clinical decisions.

Furthermore, the integration of AI into psychotherapy introduces complex questions regarding professional roles, responsibility, and patient autonomy. Scholars argue that excessive reliance on automated systems may weaken human-centered care by diminishing empathy, trust, and interpersonal communication—elements that are essential to effective therapeutic engagement (Bloch-Atefi, 2025).

3. Ethical Frameworks and Governance Approaches

In response to these concerns, researchers have proposed various ethical and governance frameworks aimed at guiding the responsible use of AI in healthcare. The foundational principles articulated by Tom L. Beauchamp and James F. Childress—autonomy, beneficence, non-maleficence, and justice—remain central to contemporary discussions, providing a normative basis for evaluating AI applications in clinical settings.

More recent approaches have expanded these principles to address the unique challenges posed by AI systems. For instance, the "ethics of care" framework emphasizes relational responsibility, emotional engagement, and the protection of vulnerable populations, offering a complementary perspective that is particularly relevant in psychotherapeutic contexts (Tavory, 2024). Additionally, global initiatives and interdisciplinary studies advocate for transparency, explainability, fairness, and accountability as key pillars of ethical AI governance (Floridi et al., 2018; Jobin et al., 2019).

Despite these developments, the literature reveals a significant gap in the integration of these ethical principles into coherent, operational frameworks tailored specifically to mental healthcare. Existing studies often remain either highly technical or purely normative, with limited efforts to bridge the gap between ethical theory and clinical implementation.

4. Research Gap and Contribution

While prior research has extensively examined both the opportunities and risks associated with AI in mental health, there remains a lack of integrative, human-centered models that systematically align technological innovation with ethical responsibility and clinical practice. Specifically, the literature lacks comprehensive frameworks that combine ethical principles, practical implementation strategies, and continuous governance mechanisms within psychotherapeutic contexts.

Addressing this gap, the present study contributes by synthesizing existing knowledge into a structured analytical framework and proposing a human-centered, ethically grounded model for the responsible integration of AI in psychotherapy.

Methodology

This study adopts a qualitative, systematic, and integrative review methodology aimed at critically analyzing the ethical implications of AI applications in mental healthcare and developing a conceptual framework for responsible implementation.

1. Research Design

The research is based on a systematic literature review combined with conceptual analysis, enabling a comprehensive synthesis of existing theoretical and empirical studies. This approach is particularly appropriate for examining complex and interdisciplinary topics where empirical data alone are insufficient to capture ethical and conceptual dimensions.

2. Data Sources and Search Strategy

Relevant literature was identified through structured searches in major academic databases, including:

- Scopus
- Web of Science
- PubMed
- PsycINFO

- Google Scholar

The search strategy employed a combination of keywords such as:

- "Artificial Intelligence in Mental Health"
- "AI Ethics in Psychotherapy"
- "Algorithmic Bias in Healthcare"
- "Digital Mental Health"
- "Explainable AI in Clinical Practice"

The search was limited to publications between 2015 and 2025 to ensure the inclusion of the most recent developments, while seminal earlier works were also incorporated where theoretically relevant.

3. Inclusion and Exclusion Criteria

The selection of studies was guided by the following criteria:

Inclusion criteria:

- Peer-reviewed journal articles and high-quality conference papers
- Studies addressing AI applications in mental health or psychotherapy
- Research focusing on ethical, legal, or clinical implications
- Publications in English

Exclusion criteria:

- Studies lacking academic rigor or methodological transparency
- Articles focusing solely on technical AI development without ethical or clinical relevance
- Non-scholarly sources and opinion pieces without empirical or theoretical grounding

4. Data Analysis

The selected literature was analyzed using a thematic and comparative approach, allowing for the identification of key patterns, recurring ethical concerns, and conceptual frameworks. The analysis focused on:

- Types of AI applications in mental healthcare
- Ethical risks and challenges
- Existing governance models and ethical principles
- Gaps between theory and practice

These elements were systematically categorized and synthesized to construct an integrated analytical framework.

5. Development of the Conceptual Framework

Based on the findings of the literature review, the study develops a conceptual model for ethical AI integration in psychotherapy, combining:

- Classical bioethical principles
- Contemporary AI ethics guidelines
- Human-centered design approaches

This framework emphasizes transparency, accountability, human oversight, and continuous ethical evaluation as core components of responsible AI deployment.

6. Limitations

While the study provides a comprehensive synthesis of current knowledge, it is limited by its reliance on secondary data and the absence of primary empirical validation. Future research is encouraged to test the proposed framework through empirical studies, case analyses, and clinical applications.

1. The Use of Artificial Intelligence in Psychotherapeutic Interventions

The use of AI in the humanities and social sciences has given rise to four major concerns, commonly referred to in scientific discourse as the "four horsemen of the AI apocalypse." This metaphor, derived from the Four Horsemen in the Book of Revelation in the Bible—symbolizing war, famine, pestilence, and death—captures a set of existential risks associated with AI development. In the context of AI, these concerns include:

- The displacement of human labor by intelligent systems,
- The potential for machines to make unethical decisions,
- Hostile or adversarial actions against humans led by autonomous systems,
- Opaque and non-interpretable "black box" decision-making processes.

Conversely, many scholars express optimism regarding the transformative potential of AI in mental health, including:

- The development of behavioral digital biomarkers,
- The redefinition of diagnostic frameworks,
- Facilitating early detection of mental disorders,
- The creation of continuous learning systems capable of assessing patients within their real-life contexts,
- Providing tools that enhance understanding of mental illness and human psychology,
- Enabling personalized diagnostic and therapeutic approaches,
- Integrating advanced computational models to improve the safety, efficiency, and personalization of mental healthcare.

This tension between exaggerated fears and ambitious expectations has led to the emergence of the concept of **Artificial Wisdom (AW)**, which represents a forward-looking paradigm. AW seeks to embed AI applications within the human context of the patient, balancing ethical and psychological considerations, thereby making it a safer and more effective tool for supporting mental health (Lee et al., 2022).

Artificial Intelligence encompasses a wide range of technologies that enable computational systems to perform human-like cognitive processes such as learning, reasoning, problem-solving, pattern recognition, generalization, and predictive inference. Broadly speaking, researchers (D'Alfonso, 2020) identify three primary domains for AI applications in mental health:

• Personal Sensing or Digital Phenotyping

Digital phenotyping, also referred to as personal sensing, involves the use of data collected from personal digital devices—particularly smartphones—to gather behavioral and contextual information about individuals. These data are subsequently used as inputs for machine learning methods to predict mental states and psychological outcomes. In addition to smartphones, wearable devices such as smartwatches and activity trackers are also employed.

• Natural Language Processing of Clinical Texts and Social Media Content

The premise that language and vocal expression reflect psychological states has driven advances in natural language processing (NLP) and speech analysis. Clinical interview transcripts serve as traditional sources for mental health-related language analysis, alongside social media content and online forums. These technologies include speech recognition, sentiment analysis, lexical and semantic analysis, and optical character recognition (OCR), all aimed at transforming unstructured text into structured data suitable for further analysis. NLP techniques are particularly relevant to psychiatry, as language and speech constitute primary sources of information in diagnosing and treating mental disorders (Lee et al., 2022, p. 859).

• Chatbots and Virtual Agents

A chatbot is a computer program designed to simulate conversation with users through text or voice interfaces, using either rule-based systems or advanced NLP techniques. The history of chatbots is closely linked to psychology. One of the earliest and most influential systems was ELIZA, which raised important philosophical and psychological questions. ELIZA enabled natural language interaction between humans and computers and was implemented on the MAC time-sharing system at the Massachusetts Institute

of Technology (MIT). It was written in the SLIP programming language within the MAD-SLIP environment for IBM 7091 computers (Weizenbaum, 1966).

Despite the significant advancements of AI in medical diagnosis, its routine adoption in mental health practice remains limited. This delay can be attributed to several factors, including the sensitive nature of data generated through patient-provider interactions—such as clinical notes and therapeutic conversations—as well as the complexity and multidimensionality of diagnostic criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). These data types and clinical decision-making processes are far more complex than well-defined and objective tasks (e.g., tumor detection from medical imaging), where current AI models have demonstrated strong performance. Furthermore, AI systems are typically "data-hungry," whereas mental health research often faces limited access to large, high-quality, and reliable datasets (Lee et al., 2022).

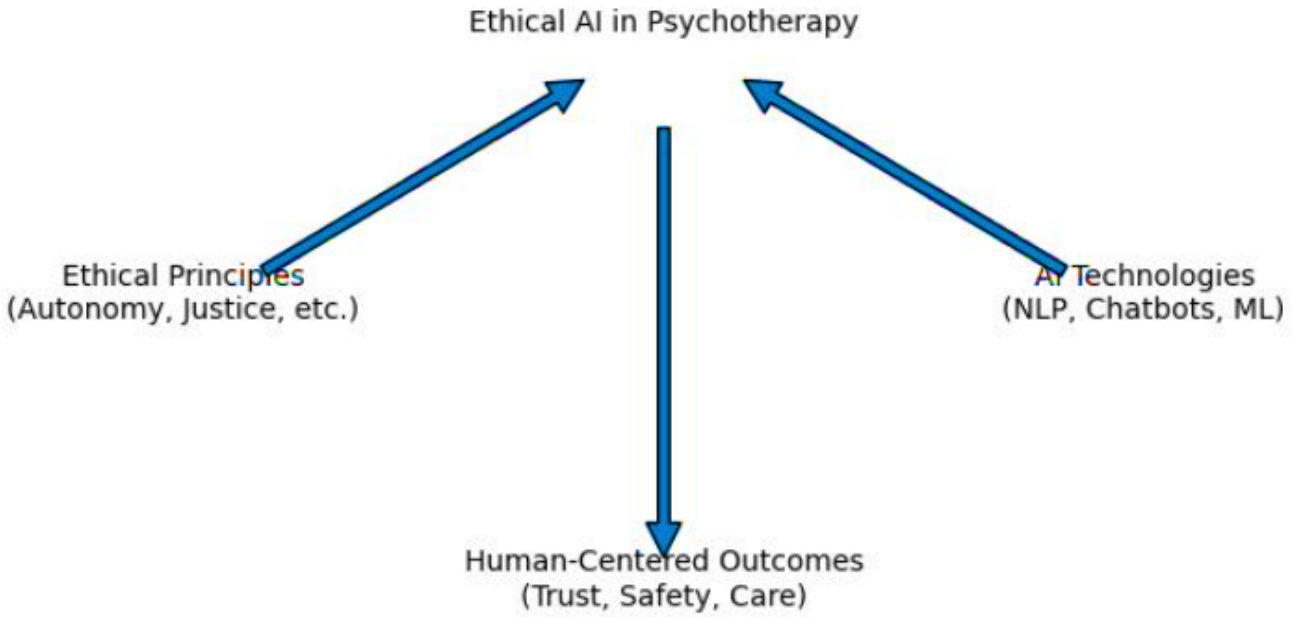

Figure 1. Conceptual Framework for Human-Centered Ethical AI Integration in Psychotherapy (Source: Author's own elaboration based on the literature)

2. Ethical Issues in the Use of AI for Diagnosis, Prediction, and Psychotherapy:

The application of AI in psychotherapy raises a number of complex ethical issues, particularly regarding the need to uphold ethical standards in its use. Numerous scholars have addressed these concerns, proposing frameworks for the ethical deployment of intelligent systems.

Fiske, Henningsen, and Buyx (2019) examined the increasing use of embodied AI applications in mental health fields, including psychiatry, psychology, and psychotherapy. These applications encompass virtual therapists and social robots used in the care of conditions such as dementia, autism, and sexual disorders. While these technologies offer promising therapeutic opportunities—such as improving access to care, enhancing patient engagement, and reducing professional workload—they also introduce significant ethical and societal challenges. These include data protection and security concerns, the absence of robust regulatory and ethical frameworks, the risk of replacing traditional healthcare services, and potential long-term impacts on our understanding of mental illness and human nature.

Warrier et al. (2023) provide a comprehensive overview of ethical challenges in AI-driven mental healthcare, including issues of privacy, informed consent, transparency, and algorithmic bias. The study offers policy recommendations, emphasizing the need for clear and globally applicable ethical guidelines governing the use of AI to enhance mental health outcomes. By balancing innovation with ethical considerations, it is possible to promote the well-being of individuals with mental disorders while safeguarding their privacy, dignity, and equitable access to care. Similarly, Pathak et al. (2024) highlight the importance of addressing ethical concerns such as bias and respect for individual privacy in the integration of AI into mental health services.

Ojo, Yewande (2024) examines ethical challenges associated with AI in psychological diagnosis and treatment planning, concluding that privacy and data protection lie at the core of ethical concerns. Psychological data require stringent safeguards to prevent unauthorized access and misuse, while still enabling responsible data use for AI-driven innovation. The study also emphasizes

fairness, non-discrimination, and the mitigation of racial bias in AI systems, as well as the need for explainability in diagnostic and therapeutic decision-making processes.

Tavory (2024) provides a comprehensive review of ethical challenges in AI-based mental healthcare, arguing that current "responsible AI" approaches are insufficient because they overlook the impact of AI on human relationships and emotional dynamics. She proposes a complementary framework based on the Ethics of Care, which emphasizes responsibility, human interaction, and the protection of vulnerable populations. This approach is particularly relevant for AI systems operating without direct human therapist involvement, ensuring protection against emotional harm, manipulation, and abrupt discontinuation of support.

A systematic review by Saeidnia et al. (2024) identifies eighteen key ethical considerations in AI-based mental healthcare, including privacy, explicit informed consent, bias mitigation, transparency, and autonomy. The study recommends involving patients' families and stakeholders in oversight processes, as well as conducting regular ethical audits and continuous monitoring of AI applications.

Beg et al. (2024), in a review of 28 studies published between 2009 and 2023, highlight the promising potential of AI-supported cognitive behavioral therapy (CBT) applications and conversational agents in alleviating symptoms of anxiety and depression. The study emphasizes the importance of cautious integration, ensuring respect for patient privacy, trust-building, and appropriate regulation of human–machine interactions.

Farmer et al. (2024) critically analyze both the opportunities and ethical/legal considerations of AI in psychological practice. While AI can reduce administrative burdens and improve service delivery, it also poses risks such as bias introduction, erosion of professional skills, and privacy concerns. The authors advocate for balanced and responsible integration, emphasizing continuous evaluation, ethical oversight, and legal compliance.

Cross et al. (2024) conducted two online surveys in Australia to assess AI usage among the general public and mental health professionals. The final sample included 107 community members and 86 professionals. While public attitudes were generally neutral, professionals showed more positive attitudes. AI usage was reported by 28% of community participants—primarily for quick support (60%) or as a personal therapist (47%)—and by 43% of professionals, mainly for research (65%) and report writing (54%). Despite overall positive perceptions of usefulness, many participants reported concerns or harms, including reduced human interaction, ethical issues, privacy risks, medical errors, and data security concerns.

Cruz-Gonzalez et al. (2025), in a systematic review of 85 studies across major databases (CCTR, CINAHL, PsycINFO, PubMed, and Scopus), identify three main domains of AI application in mental health: diagnosis, monitoring, and therapeutic intervention. Their findings indicate increasing use of AI across various disorders, with demonstrated accuracy in detection, classification, risk prediction, and treatment response forecasting. The study recommends future efforts to focus on developing more diverse and robust datasets and enhancing transparency and interpretability to improve clinical practice.

Finally, Bloch-Atefi (2025) highlights the potential of AI in counseling and psychotherapy, particularly in improving efficiency and expanding access to care in underserved areas. AI tools, such as conversational systems, can provide low-cost initial support and help reduce stigma. However, significant ethical challenges remain, including algorithmic bias, privacy violations, and limitations in establishing empathetic and trust-based therapeutic relationships. The study underscores the need to adequately train therapists—both psychologically and technically—and to develop regulatory and ethical frameworks that ensure safe, equitable, and human-centered use of AI technologies.

3. Ethical Considerations in the Use of Artificial Intelligence in Healthcare and Mental Health:

Beauchamp and Childress (2001) proposed a coherent and widely accepted ethical framework for addressing issues in healthcare, consisting of four fundamental principles that guide ethical decision-making in medical and research contexts. These principles provide a crucial foundation for analyzing the ethical dimensions of AI applications in mental health:

1. Respect for Autonomy (Beauchamp & Childress, 2001, p. 57)

This principle refers to patients' right to make informed decisions regarding their healthcare. In the context of AI, it raises several considerations, including:

- Informed consent,
- Privacy and data control,
- Shared decision-making between the patient, the system, and/or the clinician.

2. Non-maleficence (Beauchamp & Childress, 2001, p. 113)

This principle emphasizes the obligation to avoid harm, which is particularly critical in AI-supported mental healthcare. It entails:

- Minimizing potential errors,
- Addressing bias in datasets and models,
- Ensuring data protection and security.

3. Beneficence (Beauchamp & Childress, 2001, p. 165)

This principle involves actively promoting patients' well-being. In AI-based mental health applications, it includes:

- Improving quality of care,
- Expanding access to services,
- Continuous improvement of tools and interventions.

4. Justice (Beauchamp & Childress, 2001, p. 225)

This principle concerns the fair distribution of benefits and burdens. In AI contexts, it includes:

- Ensuring equitable access to services,
- Promoting algorithmic fairness.

Building on these principles, Tavory (2024) advances the **Ethics of Care** as an alternative ethical framework to traditional models grounded in responsibility or regulatory compliance. This approach is particularly relevant to AI in mental health, as it emphasizes human relationships, emotional engagement, and shared responsibility—making it well-suited to addressing the unique ethical challenges posed by AI in psychotherapy. Tavory (2024, p. 8) proposes three main domains of application:

First: Ethics of Care in the Development Phase of AI-Based Psychotherapeutic Technologies

In the absence of formal regulatory standards, it is essential to ensure the clinical validation of AI tools to confirm their safety and effectiveness. Involving patients and users in the design process is crucial for accurately understanding their psychological and social needs, alongside engaging key stakeholders such as healthcare teams and patients' families. Cultural diversity must also be carefully considered through comprehensive mapping of local communities and socio-cultural contexts.

Additionally, it is important to identify vulnerable populations and propose appropriate technological solutions tailored to their needs, recognizing that vulnerability is an inherent aspect of the human condition that may emerge at different life stages. This necessitates clear safeguards to ensure appropriate handling. Proactive mechanisms should also be developed to detect and mitigate risk factors, while integrating features that support human interaction within therapeutic system design and allow for human intervention when necessary.

Furthermore, clear strategies should be established for updating or discontinuing therapeutic AI systems (e.g., chatbots or robots), taking into account the potential psychological and emotional impacts on users.

Second: Emotional Considerations in the Use of AI in Mental Healthcare

Respect for human dignity constitutes the ethical foundation of all design and implementation processes involving intelligent systems, particularly in therapeutic contexts. It is essential to avoid exploiting users' trust or their desire for interaction, and to prevent any form of emotional manipulation.

Emotional expressions should not be treated as reliable predictors of future psychological states, requiring caution in interpretation. Emotional biases—whether affecting individuals or therapeutic relationships—must also be considered, especially given the lack of universal consensus on the definition and expression of emotions.

Moreover, cultural diversity in emotional expression and interpersonal relationships should be taken into account to ensure respectful, inclusive, and equitable interactions across different cultural backgrounds.

Third: Ethical Considerations in Using Intelligent Systems to Meet Users' Psychological Needs

Developers of AI systems in mental healthcare must commit to promoting patient well-being and supporting the therapeutic relationship where applicable, ensuring that the intended use of technology aligns with these goals. User feedback and system responses should be managed in ways that ensure actual needs are being met.

This also requires the establishment of appropriate professional and ethical policies, including risk management mechanisms—particularly for handling emergency situations requiring human intervention. Ensuring the accuracy of information and grounding it in reliable scientific sources is equally critical to prevent misinformation.

Regarding privacy, policies should go beyond minimum legal requirements by avoiding the storage of personally identifiable information whenever possible, and ensuring that such data remain confined within user-controlled applications. Data should not be shared with third parties unless legally required, and explicit, transparent user consent must always be obtained.

OjoYewande (2024) further emphasizes the need to reconsider professional ethics and responsibilities in light of the expanding role of AI in diagnosis and treatment. It is essential to clearly define the boundaries of professional judgment and legal accountability for harms that may arise from AI systems. Broader societal implications—including shifts in public perceptions of mental healthcare and transformations in the healthcare workforce—also warrant in-depth analysis.

Table 1. Ethical Risks and Governance Strategies in AI-Driven Mental Health Care

Ethical Dimension	Key Risks in AI Applications	Clinical Implications	Proposed Governance Strategies
Algorithmic Bias and Fairness	Biased training data leading to unequal diagnostic outcomes across populations	Misdiagnosis, unequal treatment recommendations, health disparities	Use diverse and representative datasets; implement bias detection and mitigation protocols; conduct regular algorithmic audits
Data Privacy and Confidentiality	Unauthorized access, data breaches, and misuse of sensitive psychological data	Loss of patient trust, legal violations, psychological harm	Apply strong encryption methods; ensure compliance with data protection regulations; adopt privacy-by-design principles
Transparency and Explainability	"Black-box" decision-making processes lacking interpretability	Reduced clinician trust, inability to justify clinical decisions	Integrate Explainable AI (XAI) models; provide interpretable outputs; ensure clinician understanding of AI recommendations
Patient Autonomy	Limited informed consent and lack of control over AI-driven decisions	Reduced patient agency and ethical violations	Implement transparent consent mechanisms; enable shared decision-making between patient, clinician, and AI system
Therapeutic Alliance	Reduced human interaction due to overreliance on AI systems	Weakening of empathy, trust, and interpersonal communication	Maintain human-in-the-loop models; ensure AI complements rather than replaces clinicians
Accountability and Responsibility	Unclear responsibility for AI-related errors or harm	Legal ambiguity and ethical uncertainty	Define clear accountability frameworks; establish regulatory oversight and clinical governance structures
Safety and Reliability	Inaccurate predictions or system failures	Clinical risks and potential harm to patients	Conduct clinical validation studies; implement continuous monitoring and evaluation systems

Table 2. Conceptual Framework for Human-Centered Ethical AI Integration in Psychotherapy

Framework Component	Description	Operational Mechanisms	Expected Outcomes
Ethical Foundations	Integration of core bioethical principles (autonomy, beneficence, non-maleficence, justice) with ethics of care	Ethical guidelines embedded in system design; interdisciplinary ethical review processes	Alignment of AI systems with human values and clinical ethics
Data Governance	Responsible collection, storage, and use of mental health data	Data anonymization, encryption, and user-controlled data access	Protection of patient privacy and data security

Algorithmic Transparency	Ensuring interpretability and explainability of AI models	Adoption of Explainable AI (XAI) techniques; transparent reporting of model decisions	Increased clinician trust and improved decision-making
Human–AI Collaboration	Hybrid model combining AI efficiency with human clinical judgment	Human-in-the-loop systems; clinician supervision of AI outputs	Enhanced accuracy while preserving human oversight
Continuous Ethical Auditing	Ongoing evaluation of AI systems for ethical compliance	Periodic audits, bias detection tools, and stakeholder feedback mechanisms	Early detection of risks and sustained ethical alignment
Stakeholder Engagement	Inclusion of patients, clinicians, and policymakers in AI development	Participatory design approaches; user-centered system testing	Improved usability, acceptance, and contextual relevance
Regulatory and Legal Frameworks	Establishment of clear policies governing AI use in healthcare	Development of adaptive regulations; compliance with international standards	Legal clarity, accountability, and safe deployment of AI systems

Regulatory frameworks and governance mechanisms play a critical role in addressing these challenges. Policymakers face the complex task of developing flexible regulations that foster innovation while ensuring robust ethical safeguards. This requires a collaborative, interdisciplinary approach involving clinicians, psychologists, ethicists, researchers, engineers, and AI developers.

AI has the potential to transform mental healthcare by improving diagnostic accuracy, enabling personalized treatment, and enhancing therapeutic outcomes. It also contributes to increased efficiency, affordability, and accessibility of care. Tools such as chatbots, virtual therapists, and predictive algorithms are already emerging. However, ethical guidelines and responsible practices remain essential to ensure that AI contributes positively to the well-being of individuals with mental disorders (Warrier, Warrier, &Khandelwal, 2023).

Proposed Ethical Considerations for Future AI-Based Psychotherapy Applications:

1. Algorithmic bias represents a major concern in the diagnosis and treatment of mental disorders. AI systems rely on large datasets that may contain inherent biases, leading to disparities in diagnosis and treatment recommendations, particularly affecting marginalized populations.

2. Privacy and data protection are among the most critical ethical challenges in AI-driven mental healthcare. These include risks of unauthorized access, data breaches, and the commercial exploitation of patient information, necessitating strict protective measures.

3. Maintaining ethical standards in AI-based mental healthcare is essential. The opacity of AI systems may hinder understanding of decision-making processes. Responsible use requires explainability, as well as accountability for outcomes, particularly in cases of errors or harm.

4. Transformation of the therapeutic relationship: The integration of AI may alter traditional therapist–patient dynamics by introducing advanced technological tools. Achieving a balanced integration between AI support and professional expertise remains a complex and evolving ethical challenge.

5. Explicit informed consent is a cornerstone of medical ethics. Patients must retain the right to make fully informed decisions and to refuse AI-based interventions if they have reservations.

A comprehensive study by Saeidnia et al. (2024), based on a systematic review of literature from major databases (PubMed, PsycINFO, Web of Science, and Scopus) covering the period 2014–2024, analyzed 51 relevant articles and identified 18 key ethical considerations, categorized as follows:

1. Six ethical considerations in AI-based mental health interventions:

- Privacy and confidentiality
- Informed consent
- Bias and fairness

- Transparency and accountability
- Autonomy and human agency
- Safety and effectiveness

2. Five ethical principles for development and implementation:

- Ethical frameworks
- Stakeholder engagement
- Ethical review
- Bias mitigation
- Continuous evaluation and improvement

3. Seven best practices and recommendations:

- Adherence to ethical guidelines
- Ensuring transparency
- Prioritizing data privacy and security
- Mitigating bias and ensuring fairness
- Stakeholder involvement
- Regular ethical audits
- Monitoring and evaluating outcomes

This systematic review underscores the importance of embedding ethical considerations into the responsible implementation of AI technologies in mental healthcare, ensuring the protection of individuals' dignity and the achievement of equitable and sustainable therapeutic outcomes.

Conclusion:

The integration of Artificial Intelligence into Mental Health care represents a paradigmatic transformation in the delivery, accessibility, and personalization of psychotherapeutic services. However, as this study has demonstrated, the expansion of AI-driven applications in psychotherapy cannot be understood solely through the lens of technological advancement; rather, it must be critically evaluated within a robust ethical, clinical, and regulatory framework that safeguards fundamental human values. The findings underscore that without the systematic incorporation of ethical governance mechanisms, AI technologies risk reinforcing existing inequalities, compromising patient autonomy, and undermining the therapeutic alliance that constitutes the foundation of effective mental health care.

This paper contributes to the emerging interdisciplinary discourse by advancing a human-centered and ethically grounded perspective on AI integration, emphasizing the necessity of aligning algorithmic efficiency with principles of autonomy, beneficence, non-maleficence, and justice, alongside relational approaches such as the ethics of care. In this regard, AI should not be conceptualized as a substitute for human clinicians, but rather as a complementary and augmentative tool operating within a hybrid decision-making ecosystem that preserves professional accountability and human oversight.

Despite recent efforts by certain national and international actors to establish preliminary regulatory responses, the current global landscape remains characterized by significant fragmentation and the absence of comprehensive legal frameworks governing AI in healthcare. This regulatory asymmetry creates a critical gap between technological innovation and institutional readiness, thereby exposing patients and practitioners to ethical, legal, and operational risks. Addressing this gap requires the development of adaptive, transparent, and enforceable governance architectures capable of responding to the dynamic nature of AI systems.

Accordingly, future progress in this domain depends on sustained interdisciplinary collaboration among AI developers, clinicians, ethicists, legal scholars, and policymakers. Such collaboration is essential for designing clinically validated, ethically aligned, and socially responsible AI systems that are responsive to diverse cultural and contextual realities. Furthermore, continuous ethical auditing, stakeholder engagement, and the integration of explainable AI principles are necessary to ensure accountability and trust in AI-assisted mental healthcare. In conclusion, the responsible and sustainable deployment of AI in psychotherapy necessitates a shift from purely innovation-driven approaches toward ethically informed and human-centered paradigms. Only through the

establishment of comprehensive regulatory frameworks, the reinforcement of professional oversight, and the preservation of the intrinsically human dimensions of care—such as empathy, trust, and relational engagement—can AI fulfill its potential as a transformative yet ethically responsible force in the future of mental health care.

Declarations

Ethics Approval and Consent to Participate. This study is based exclusively on previously published literature and does not involve human participants, animals, or primary data collection. Therefore, ethical approval and informed consent were not required in accordance with institutional and international research ethics guidelines.

Consent for Publication. Not applicable. This manuscript does not contain any individual person's data in any form.

Availability of Data and Materials. All data supporting the findings of this study are derived from publicly available sources cited within the reference list. No new datasets were generated or analyzed during the current study.

Conflict of Interest Statement. The authors declare that they have no competing interests, financial or non-financial, that could have influenced the work reported in this paper.

Funding

The authors received no specific funding for this work from any public, commercial, or not-for-profit funding agencies.

Authors' Contributions

- Senoussaoui, Abderrahmen: Conceptualization, literature review, methodology design, writing – original draft, and theoretical framework development.
- Rahmani, Djamel: Supervision, validation, critical revision of the manuscript, and contribution to ethical analysis and interpretation.

All authors have read and approved the final manuscript.

Acknowledgements. The authors would like to express their appreciation to the academic and research communities whose published work contributed to the development of this study.

Ethical Considerations. This study adheres to internationally recognized ethical standards in research and publication. The analysis was conducted in accordance with the ethical principles of Tom L. Beauchamp and James F. Childress, including respect for autonomy, beneficence, non-maleficence, and justice. Special attention was given to issues related to data privacy, algorithmic bias, transparency, and the preservation of human dignity in the context of Artificial Intelligence applications in Mental Health care.

AI Use Disclosure Statement. The authors confirm that no generative artificial intelligence tools were used in the creation, analysis, or writing of this manuscript in a manner that replaces human intellectual contribution. Any use of digital tools was limited to language editing and formatting assistance, under full human supervision.

Data Privacy Statement. This study does not involve the collection, processing, or storage of personal data. All referenced materials are publicly available and used in accordance with academic and ethical standards.

References

1. Beg, M., Verma, M., Verma, M., & Chanthar, V. (2024). Artificial intelligence for psychotherapy: A review of the current state and future directions. *Indian Journal of Psychological Medicine*. https://doi.org/10.1177/025371762412608
2. Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford University Press.
3. Bloch-Atefi, A. (2025). Balancing ethics and opportunities: The role of AI in psychotherapy and counselling. *Psychotherapy and Counselling Journal of Australia*. https://doi.org/10.59158/001c.129884
4. Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., Baker, S., & Alvarez-Jimenez, M. (2024). Use of AI in mental health care: Community and mental health professionals survey. *JMIR Mental Health, 11*, e60589. https://doi.org/10.2196/60589
5. Cruz-Gonzalez, P., He, A. W. J., Lam, E., Ng, I. C., Li, M., Hou, R., & Miller, T. (2025). Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine, 55*(e18), 1–52. https://doi.org/10.1017/S0033291724003295

6. D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology, 36*, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005
7. Farmer, R., Lockwood, A., Goforth, A., & Thomas, C. (2024). Artificial intelligence in practice: Opportunities, challenges, and ethical considerations. *Professional Psychology: Research and Practice, 59*(1), 19–32. https://doi.org/10.1037/pro0000595
8. Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), e13216. https://doi.org/10.2196/13216
9. Kasri, N. (2024). Regulatory principles governing the use of artificial intelligence. *Al-Turath Journal, 14*(4), 35–44.
10. Lee, E. E., Torous, J., De Choudhury, M., Depp, C., Graham, S., Kim, H.-C., & Jeste, D. V. (2022). Artificial intelligence for mental healthcare: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 7*(9), 856–864. https://doi.org/10.1016/j.bpsc.2021.02.001
11. Ojo, Y. (2024). Ethical considerations in using AI for mental health diagnosis and treatment planning: A scoping review. In *Proceedings of the International Conference on Artificial Intelligence and Robotics* (pp. 169–180). MIRG-ICAIR.
12. Pathak, H., Sood, S., & Anshul, K. (2024). AI in mental health: A comprehensive review, comparative analysis, and ethical considerations for advancing assessment and treatment. *International Research Journal of Modernization in Engineering Technology and Science, 6*(7), 583–591. https://doi.org/10.56726/IRJMETS59822
13. Saeidnia, H., Fotami, S., Lund, B., & Ghiasi, N. (2024). Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences, 13*(7), 381. https://doi.org/10.3390/socsci13070381
14. Tavory, T. (2024). Regulating AI in mental health: Ethics of care perspective. *JMIR Mental Health, 11*, e58493. https://doi.org/10.2196/58493
15. Trusler, K., Doherty, C., Mullin, T., Grant, S., & McBride, J. (2006). Waiting times for primary care psychological therapy and counselling services. *Counselling and Psychotherapy Research, 6*(1), 23–32. https://doi.org/10.1080/14733140600581358
16. Van Dijk, D., Meijer, R., van den Boogaard, T., Spijker, J., Ruhé, H. G., & Peeters, F. P. M. L. (2023). Worse off by waiting for treatment? The impact of waiting time on clinical course and treatment outcome for depression in routine care. *Journal of Affective Disorders, 322*, 205–211. https://doi.org/10.1016/j.jad.2022.11.011
17. Warrier, U., Warrier, A., & Khandelwal, K. (2023). Ethical considerations in the use of artificial intelligence in mental health. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery, 59*(1), 139. https://doi.org/10.1186/s41983-023-00735-2
18. 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
19. Zerari, M. (2025). Artificial intelligence and the right to privacy on social media from an international perspective. *Journal of Law, Society and Authority, 14*(1), 177–197.
20. Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.
21. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181
22. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259
23. Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
24. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1
25. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2
26. Torous, J., & Roberts, L. W. (2017). Needed innovation in digital health and smartphone applications for mental health. *JAMA Psychiatry, 74*(5), 437–438. https://doi.org/10.1001/jamapsychiatry.2017.0262
27. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689
28. Mittelstadt, B. D., et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679
29. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
30. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.