BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2025, Volume 16, Issue 4, pages: 424-435
Submitted: September 24th, 2025 | *Accepted for publication:* October 10th, 2025

Responsible AI Integration Framework for Psychotherapy

Shlomo Mendlovic
MSc PhD MD MHA
Faculty of Medicine, Tel Aviv University, Tel Aviv, Shalvata Mental Health Centre, Hod Hasharon, & ALUMOT: Knowledge, Creativity, Leadership in Psychotherapy - The Clalit Program for Psychotherapy, Israel.
shlomo.mendlovic@gmail.com,
https://orcid.org/0000-0002-0587-1169

Abstract: *The article explores the opportunities and risks associated with the integration of artificial intelligence (AI) into psychotherapy. AI can expand access to care, provide clinical decision support, enable continuous monitoring, and increase system efficiency. At the same time, significant challenges remain, including technological immaturity, data privacy and ethical concerns, and the potential erosion of the therapeutic alliance. To ensure safe and responsible adoption, the article introduces the Responsible AI Integration Framework for Psychotherapy, which identifies four essential capabilities: Explainable AI (XAI), Shared Decision Making (SDM), Electronic Health Record (EHR) integration, and Harm Prevention (HP). Together, these components form a practical, workflow-embedded formula that addresses both technical and ethical requirements. A roadmap for implementation is also proposed, emphasizing phased adoption within public health systems through preparation and vision, controlled pilots, gradual scaling, and full integration with ongoing improvement. The framework underscores that AI should function as a supportive tool, enhancing rather than replacing the empathic human presence that remains central to psychotherapy.*

Keywords: *artificial Intelligence; psychotherapy; responsible AI; explainable AI (XAI); shared decision making (SDM); electronic health records (EHR); harm Prevention (HP); mental health; ethical framework; public health systems; clinical decision support; digital mental health.*

How to cite: Mendlovic, S. (2025). Responsible AI integration framework for psychotherapy. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *16*(4), 424-435. https://doi.org/10.70594/brain/16.4/26

1. Introduction

Artificial intelligence (AI) is increasingly entering fields that were once thought to require only human expertise. Psychotherapy, a practice grounded in nuanced understanding, relational depth, and ethical sensitivity, is now confronted with technologies capable of naturalistic dialogue, recognizing emotions, and adaptive recommendation. However, despite a growing body of ethical discussion, the field still lacks a ready-to-use framework for deployment: existing approaches address explainability, patient autonomy, data governance, and harm prevention in isolation, leaving clinicians and policymakers without a comprehensive guide for safe adoption. Existing frameworks (e.g., IEACP, BPES) remain largely theoretical and pre-deployment; and none offer a model that integrates responsibility into workflow.

This article introduces the Responsible AI Integration Framework for Psychotherapy, which is the first operational model to define four non-negotiable system capabilities for AI-assisted psychotherapy: Explainable AI (XAI), Shared Decision Making (SDM), Electronic Health Record (EHR) integration, and Harm Prevention (HP). Moving beyond normative reflection, the framework adopts a responsibility-by-design perspective, translating abstract ethical goals into specific, workflow-embedded requirements that consider the practical realities of public mental health systems.

Our approach is deliberately pragmatic and aware of policy implications. We conduct a systematic SWOT analysis to clarify the operational risks associated with opaque, non-collaborative, or poorly supervised systems. We then translate these insights into a single, actionable formula to guide developers, regulators, and health systems in making AI both clinically trustworthy and ethically defensible. Throughout this process, we emphasize that AI should act as a support system, enhancing, rather than diminishing, the empathic human presence at the heart of psychotherapy.

2. Integrating Artificial Intelligence into Psychotherapy: Strengths, Weaknesses, Opportunities, and Threats

Artificial intelligence (AI) is rapidly transforming mental health care, particularly in psychotherapeutic practice (Gual-Montolio et al., 2022, Gutierrez et al., 2024, Dehbozorgi et al., 2025, Sun et al., 2025). From natural language processing systems that analyse session transcripts to machine-learning models that predict treatment response, AI technologies are being explored as adjuncts to human clinicians. However, the prospect of embedding AI within psychotherapy is complex and multifaceted. A structured SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis clarifies both the potential value and the risks associated with technological transition.

2.1. Strengths

One of the most frequently cited strengths of AI integration is the potential to **improve access to care** (Dehbozorgi et al., 2025). Digital systems can provide screening, triage, and psychoeducational support to individuals in remote or underserved regions, reducing barriers such as geographic distance and limited clinician availability. AI can also provide **clinical decision support** (e.g., Benrimoh et al., 2025, see Auf et al., 2025 for review), by analysing session transcripts, identifying patterns in speech or affect, and suggesting tailored interventions that complement, rather than replace, the therapist's judgment. Such tools can serve as a "second set of eyes," helping clinicians detect risk markers such as suicidality or symptom deterioration.

Another advantage is its ability for **continuous measurement and evaluation** (Cruz-Gonzalez et al., 2025). AI systems can automatically capture and analyse large quantities of data from therapy sessions, patient self-reports, and wearable sensors, supporting outcome tracking and quality improvement. This monitoring may help clinicians adapt treatment in real time. Moreover, AI can promote **system-level efficiency** (Stephenson et al., 2025) by enabling intelligent triage, automated follow-up, and stepped or stratified care models. These approaches allocate

resources dynamically according to patient need and response, potentially reducing waitlists and minimizing clinician burnout.

Finally, AI offers avenues for **new professional competencies**. Systems that provide feedback on therapist interventions or simulate challenging patient scenarios can support deliberate practice and skill development (Mendlovic, 2025). Such technologies may enrich training programs, making supervision more data-driven and scalable.

2.2. Weaknesses

Despite these strengths, current AI tools often exhibit **technological and clinical immaturity** (Graham et al., 2019). Many systems struggle to interpret nuanced emotional processes, implicit meanings, or complex interpersonal dynamics that define psychotherapy. Overreliance on shallow text analysis or pattern recognition is a potential risk, because missing the deeper, contextualised understanding essential to therapeutic work.

A second concern is the **potential erosion of the human connection** (Grodniewicz, 2023, Frances, 2025). The therapeutic alliance—marked by trust, empathy, and attunement—remains a central predictor of positive outcome. If AI systems are introduced insensitively, they may undermine the relational fabric of therapy, reducing patients' sense of being heard and understood. Relatedly, **data security and ethical uncertainty** (Mohsin Khan et al., 2025) present substantial vulnerabilities. Handling highly sensitive mental health information demands robust privacy protections and clear regulatory frameworks. Clearly, current standards remain insufficient to assure protection and safety. AI models are also **data-dependent**: a well done processing requires large, diverse, high-quality clinical datasets. Unfortunately, such datasets are often scarce, fragmented, and at times biased (Monteith et al., 2022).

This limits generalisability and may perpetuate inequities. Finally, evidence regarding the **long-term effectiveness** is still emerging (Humayun et al., 2025) as few rigorous randomised controlled trials have demonstrated sustained improvement in real-world psychotherapy outcomes.

2.3. Opportunities

Looking ahead, several opportunities can enhance the responsible deployment of AI in mental health care. As further detailed, adopting explainable and comprehensible systems that preserve the human element could make recommendations more transparent for both therapist and patient, thereby **preserving autonomy and clinical accountability**. Second, developers can build **culturally and linguistically tailored models** that are sensitive to local norms and diverse populations (see, for example, Grenon et al., 2025). Third, **integration with electronic health records** (EHRs) could improve coordination across multidisciplinary teams and facilitate continuous care monitoring (Hawley, 2021). Importantly, AI can be positioned as a **complement rather than a replacement** (Kellogg & Sadeh-Sharvit, 2022), handling screening, routine monitoring, and skill rehearsal clinicians can focus on empathy and complex formulation. Finally, the field is experiencing a surge of **innovation and investment** with start-ups, venture capital, and health regulators exploring new business models and infrastructures to responsibly scale AI-enhanced psychotherapy (see, for example, (https://www.grandviewresearch.com/industry-analysis/ai-mental-health-market-report).

2.4. Threats

Nevertheless, significant threats must be addressed. **Regulatory uncertainty** remains a critical challenge: evolving privacy laws, liability standards, and risk-management rules create ambiguity for clinicians and developers (El Arab et al., 2025). There is a **risk of patient harm** if prediction errors, biased outputs, or poorly supervised advice go unchecked (Tekin & Delehanty, 2025). **Professional resistance** is also possible (Arvai, 2025); some therapists worry that AI could undermine their expertise or lead to deskilling, slowing adoption. **Trust** may be compromised by unethical use, data breaches, or security failures, all of which could delegitimise the entire field

(Mohsin Khan et al., 2025). Finally, the pace of AI advancement may **outstrip clinicians' ability to train and adapt**, leaving many feeling unprepared or disempowered. Patients, who may at times be in an anxious psychological state, might also feel "abandoned to AI" by their healthcare providers, which could lead to heightened risk-averse and conservative behavior regarding their medical care (Gundlack et al., 2025).

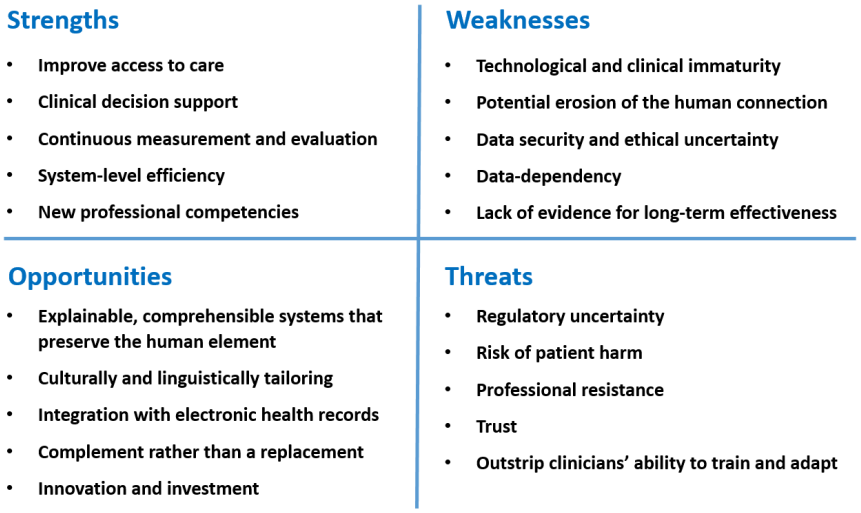

Figure 1. SWOT analysis of AI-integration in psychotherapy

AI has the potential to enhance psychotherapy by increasing access, deepening insight, and supporting data-driven practice. However, it also presents technical, ethical, and relational challenges that must be addressed. To maximize its benefits, careful regulation, robust evidence, explainable and culturally sensitive design, and a collaborative approach that empowers —not replaces— clinicians are essential.. A clear understanding of strengths, weaknesses, opportunities, and threats can guide responsible innovation and sustain the core human values of psychotherapy in an AI-augmented era.

3. Artificial Intelligence in Public Psychotherapy Systems: Unique Advantages and Strategic Potential

Integrating artificial intelligence (AI) into psychotherapy offers significant advantages for public mental health systems, particularly those facing long waitlists, limited resources, and the responsibility to serve diverse populations. Unlike private care, public systems must balance efficiency, equity, and clinical quality at scale, a balance AI can help achieve.

Expanding access and equity: AI tools, such as triage systems, online scheduling and conversational agents can streamline intake processes, shorten waiting times, and provide psychoeducation or crisis screening. These tools help reach underserved populations across different languages and geography, reduce travel barriers and enable earlier intervention.

Leveraging rich data: Public systems maintain unified electronic health records (EHRs), that capture longitudinal treatment histories and population trends. This centralised data can power predictive models for risk stratification, treatment personalisation, and resource allocation- advantages fragmented private data lacks.

Driving efficiency and cost control: AI-enabled monitoring, early detection of patient deterioration and stepped or stratified care can reduce unnecessary hospitalisations and free clinician time. For budget-constrained systems, this means shorter waits, better utilization of workforce, and stronger crisis response capabilities.

Standardising and safeguarding quality: Public agencies (as opposed to fragmented private care) can embed AI into existing digital infrastructures, enforce evidence-based protocols and monitor fairness.

Shaping responsible adoption: Governments can issue explainability and safety guidelines, negotiate with vendors for privacy and accountability, and collaborate with start-ups and academic institutions to innovate under public-interest governance.

Public mental health systems' scale, data richness, regulatory authority, and equity mandate make them ideal for AI-enhanced psychotherapy if they invest in clinician training, infrastructure, and ethical safeguards to overcome bureaucratic inertia.

4. A Framework for the Responsible Integration of AI into Psychotherapy

To facilitate the responsible integration of AI systems into psychotherapy (that is, integration which takes into account the weaknesses and risks inherent in such implementation), the following formula is proposed:

XAI + SDM + EHR + HP = Psychologist – (human factor)

This formula integrates various components which, through their combination, address the risks associated with using AI in psychotherapy.

4.1. Explainable AI (XAI)

XAI refers to artificial intelligence systems that provide clear and understandable explanations for their actions and decisions (European Data Protection Supervisor, 2023).

Explainability has crucial benefits across several domains (Amann et al., 2020): Technological (accelerating the development of more effective AI systems), legal (strengthening patients' ability to provide informed consent for treatments, facilitating regulatory approval and medical device certification of AI systems, and clarifying legal accountability when using AI in mental health) and medical-ethical (upholding central ethical values in medicine by promoting transparency and responsibility).

As a result, there is an increasing demand for explainable AI approaches in psychiatric diagnosis and prediction (Kerz et al., 2023).

Combi et al. (2022) proposed a conceptual and functional definition of XAI in medicine: an AI system able to provide clear, accessible, and contextually relevant explanations so that users can understand, evaluate, and trust its outputs in medical decision-making. They emphasise that explanations should be context-aware, tailored to the target audience (patients, clinicians, regulators) and aligned with the purpose of use, while maintaining transparency, interpretability and accountability. According to Combi et al. (2022), a medical XAI system should meet seven essential conditions:

1. Human agency and oversight: Users must maintain control over the system's decision-making processes, including the ability to intervene, provide feedback, and stop usage when necessary.
2. Technical robustness and safety: The system must be reliable, resilient to failure, and prioritize patient safety throughout its use, while identifying and managing technological risks.
3. Privacy and data governance: It must protect patient confidentiality, manage data ethically and securely, and comply with regulatory requirements.
4. Transparency: Clear and accessible explanations of how the system operates, including its data sources, and the logic behind its predictions or recommendations.
5. Diversity, non-discrimination, and fairness: Ensuring the system does not discriminate, is fair, and is built on diverse datasets representing the population.
6. Societal and environmental well-being: Evaluating the system's social and environmental impact to ensure that it promotes public welfare and avoids causing harm.
7. Accountability: Clear and practical responsibility throughout the stages of development, deployment, and user interaction.

Explainable AI can therefore mitigate the disadvantages of "black box" AI by improving transparency and trust among users (Markus et al., 2020). Explainability may be model-based (e.g., using interpretable models that can identify the most influential factors behind decisions), post-hoc (applying techniques to explain black-box models), global (offering explanations for the model as a whole), or local (providing explanations for individual predictions). Together, these approaches promote user understandability ensuring that the system is transparent and that users can interpret how it operates.

4.2. Shared Decision Making (SDM)

Shared Decision Making (SDM) in mental health refers to a collaborative process where clinicians and patients work together to make treatment decisions together (Slade, 2017).

Typically, SDM includes three key aspects: sharing relevant information, discussing available treatment options, and reaching a mutual agrement on the treatment plan. SDM offers many benefits, including increased patient satisfaction, improved adherence to both pharmacological and psychotherapeutic interventions, enhanced quality of life and reduced rates of re-hospitalisation. More importantly, SDM empowers patients by strengthening their autonomy and control, improving therapeutic alliance, and fostering greater trust between patients and their healthcare teams. It enables personalisation of treatment according to patient values and preferences, contributing to a reduction of cultural and ethnic gaps in care and supporting authentic partnership among patients, families, and multidisciplinary teams (Chmielowska et al., 2023).

At the same time, implementing SDM presents challenges, such as self-stigma and societal stigma that undermine patients' ability to participate, workforce shortages and time constraints, a high symptom burden, and skepticism about one's ability to influence medical decisions (Mertens et al., 2025).

4.3. Electronic Health Record (EHR)

Electronic Health Records (EHRs) address several key issues that are problematic in AI-supported psychotherapy, particularly regarding data quality, continuity of care, and ethical oversight. EHRs provide a standardised, longitudinal, and structured repository of patient information, which is essential for training and validating AI models in psychotherapy. This helps mitigate the risk of fragmented or incomplete data that could undermine the reliability and generalisability of AI-driven interventions (Edgcomb & Zima, 2019, Grzenda & Widge, 2023).

EHRs also facilitate the integration of patient-reported outcomes and clinician notes, supporting more robust and transparent machine learning approaches for treatment prediction and monitoring. This helps overcome methodological flaws and deficiencies in transparency that are common in current AI applications for mental health (Grzenda & Widge, 2023). Moreover, EHRs enhance accountability and medico-legal documentation, which is critical given the ethical challenges of AI in psychotherapy, such as privacy, confidentiality, and the need for clinician oversight (Aggarwal & Bhugra, 2025). Overall, by enabling secure data storage and standardised documentation, EHRs help ensure that AI-supported psychotherapy adheres to established clinical and ethical standards, thereby reducing risks of bias, loss of clinical judgment, and breaches of patient confidentiality (Aggarwal & Bhugra, 2025).

4.4. Harm Prevention (HP)

Current evidence-based risk monitoring for AI-driven psychotherapy combines technical safeguards, staged evaluations, ethical oversight, and post-market surveillance to prevent harm.

Technical guardrails automatically detect and manage harmful inputs (suicidal ideation, psychosis) as well as unsafe outputs (inappropriate advice, diagnosis). These guardrails undergo validation through both bench testing and post-deployment monitoring, with RCTs showing prevention of unsafe responses and no serious adverse events (Campellone et al., 2025). The Biological-Psychological, Economic, and Social (BPES) (Khan &, Seto, 2023) framework

incorporates safety checks across biological, psychological, economic, and social domains, using phased trials and stakeholder input to ensure “do no harm". However, A study conducted in 2025 systematically evaluated 29 commercially available AI-powered mental health chatbot agents. The findings revealed that none of the chatbots provided adequate responses to simulated suicidal risk scenarios, with nearly half failing to offer even marginally appropriate guidance such as emergency contact information or crisis referral (Pichowicz et al., 2025).

XAI + SDM + EHR + HP

To advance from abstract ethical reflections to practical application, we propose the **Responsible AI Integration Framework for Psychotherapy**. This framework defines four non-negotiable system capabilities — Explainable AI (XAI), Shared Decision Making (SDM), Electronic Health Record (EHR) Integration, and Harm Prevention (HP) — and provides an operational checklist for evaluating whether an AI system is safe and meaningful for psychotherapeutic use. Table 1 summaries the suggested operational checklist.

Table 1. Responsible AI Integration Framework for Psychotherapy

System capabilities	Providing	Minimum requirement	Indicators to verify
Explainable AI (XAI)	Transparency & interpretability	The system must provide clinically intelligible explanations for both model logic and individual recommendations	● Clinicians can review why a specific recommendation was made (local interpretability) ● A global description of model inputs, training data sources, and limitations is accessible ● Explanation complexity is matched to the intended user (patient, clinician, regulator)
Shared Decision Making (SDM)	Preserving autonomy	AI outputs function as inputs to collaborative discussion, not covert determinants of care	● Recommendations are presented with options, not imperatives ● User interface supports clinician–patient dialogue (e.g., editable suggestions, pros/cons) ● Consent and patient preferences are explicitly logged when AI informs treatment choice
EHR Integration	Clinical continuity & accountability	The AI tool must operate within the patient’s longitudinal clinical record to ensure context, oversight, and documentation	● AI pulls and writes structured notes into the EHR ● Predictions are contextualized by prior diagnoses, risk flags, and treatment history ● Audit trails allow clinicians, auditors, and regulators to reconstruct AI-influenced decisions
Harm Prevention (HP)	Multilayered safety controls	Built-in guardrails detect and respond to risk (suicidality, psychosis, harmful advice) with human-in-the-loop escalation	● Pre-deployment stress testing with high-risk scenarios ● Continuous post-deployment monitoring and override options ● Crisis pathways (e.g., emergency referrals, safety alerts) are explicit and reliable.

5. A Practical Roadmap for Integrating AI into Psychotherapy in Public Health Systems

The successful adoption of artificial intelligence (AI) in public psychotherapy services requires more than just technological readiness; it demands strategic planning, ethical oversight,

and careful change management. The following roadmap outlines a phased approach for large-scale, responsible implementation within national or regional mental health systems.

5.1. Preparation and Vision

The first step is to establish a clear strategic and ethical foundation before introducing AI tools. Health authorities and mental health leaders should define concrete goals, such as reducing wait times, improving therapy quality and supervision, and enabling early detection of high-risk situations (e.g., suicidality, severe depression).

A multidisciplinary steering committee—including psychiatrists, psychotherapists, ethicists, data scientists, cybersecurity experts, and patient representatives—should guide this stage. This committee can set priorities, identify risks, and ensure alignment with both international regulations (. GDPR and HIPAA) and local legal frameworks. Ethical principles such as Explainable AI (XAI), Shared Decision Making (SDM) and Harm Prevention (HP) should be integrated from the outset, ensuring transparency and preserving patient autonomy.

At the same time, public providers should map existing data assets. Auditing electronic health records (EHRs) and psychiatric data repositories allows for the identification of quality gaps (such as unstructured notes or insufficient labeling of emotional content) that must be addressed for effective AI training.

5.2. Controlled Pilots (Proof of Concept)

The second phase focuses on testing AI tools in well-defined, clinically supervised environments. Public systems should select focused use cases, such as automated screening and triage, conversational analysis for monitoring the therapeutic alliance, or AI-assisted deliberate practice for therapist training.

Partnerships with start-ups and academic research groups can accelerate development of culturally and linguistically adapted models. Pilot projects should apply objective evaluation metrics, including changes in wait times, patient and clinician satisfaction, and the accuracy of risk detection. Early, evidence-based feedback is essential for building trust among stakeholders and facilitating broader adoption.

5.3. Gradual Scaling

Once pilot data demonstrate safety and efficacy, public systems can scale AI solutions progressively. A crucial step in this process is seamless integration with existing infrastructures, particularly EHR platforms and stepped or stratified care pathways.

Equally important is building workforce capacity. Public mental health professionals should undergo multi-level training that ranges from basic AI literacy to advanced use of decision-support dashboards in daily clinical work. Specialised fellowship tracks or continuing education programs in digital mental health can prepare future leaders.

At this stage, a dedicated quality assurance (QA) unit should monitor clinical performance, identify algorithmic bias, and establish feedback channels for therapists and patients.

5.4. Full Implementation and Continuous Improvement

In the final stage, AI becomes a routine, trusted, and transparent component of care. Systems should adopt a human–machine collaborative model, where AI provides data-driven insights but therapists maintain ultimate clinical judgment. Patients should be informed about AI involvement, supporting shared understanding and preserving the therapeutic alliance.

Data collected across the system can power a Learning Health System, that continuously improves models based on real-world outcomes. At the same time, privacy frameworks, ethical standards, and clinical guidelines must develop alongside technological advances. Ongoing clinical research, including randomised controlled trials (RCTs), is essential to validate long-term therapeutic benefits.

When implemented thoughtfully, AI can act as a force multiplier for public mental health services by improving access, increasing quality, and optimising resources. Success depends on gradual adoption, strong ethical governance, workforce training and transparent communication. Instead of replacing therapists, AI should empower and extend their capabilities, preserving the human essence of psychotherapy while enabling systems to meet growing demand at scale.

6. Discussion

Artificial intelligence (AI) is increasingly entering fields traditionally seen as inherently human, such as psychotherapy.

Recent advancements, particularly in large language models capable of extended dialogue, affect detection, and provide context-sensitive recommendations, have sparked optimism about enhancing mental health care through technology.. However these have also exposed a significant conceptual gap: although the literature increasingly discusses explainability, patient autonomy, data governance, and harm prevention, these considerations are typically addressed in isolation. What remains missing is a coherent, operational framework that specifies the minimum structural and ethical requirements that an AI system must meet before it can be responsibly integrated into psychotherapeutic practice.

This article aims to fill that gap by proposing an integrated, practice-oriented model that defines four foundational components: Explainable AI (XAI), Shared Decision Making (SDM), Electronic Health Record (EHR) integration, and Harm Prevention (HP). These components are prerequisites for any AI aspiring to participate in psychotherapy. To our knowledge, this is the first attempt to synthesise these elements into a single actionable framework, explicitly designed to address the dominant barriers to safe and meaningful AI integration: algorithmic opacity and mistrust, erosion of patient autonomy, fragmentation of clinical data, and the risk of harm from unsupervised or biased outputs.

Despite the numerous ethical and design-oriented frameworks for AI in mental health have been proposed, none fully integrate the practical and structural requirements necessary for safe clinical deployment. The Integrated Ethical Approach for Computational Psychiatry (IEACP) (Putica et al., 2025), for example, provides a valuable conceptual framework for responsible decision-making but remains largely normative and abstract, without specifying concrete technological or workflow prerequisites. Similarly, the Biological-Psychological-Economic-Social (BPES) safety checklist (Khan & Seto, 2023) provides an important risk–benefit evaluation tool but it focuses more on pre-deployment feasibility assessment rather than ongoing clinical integration. In contrast, the model presented in this article combines Explainable AI (XAI), Shared Decision Making (SDM), Electronic Health Record (EHR) integration, and Harm Prevention (HP) into a single, actionable framework to designed for real-world implementation within psychotherapeutic practice. This synthesis moves beyond high-level ethical aspiration by identifying non-negotiable system capabilities, aligning them with routine clinical workflows, and emphasising the unique opportunities and constraints of public mental health infrastructures. To our knowledge, this is the first attempt to consolidate these four domains into a pragmatic, deployment-ready blueprint for AI-assisted psychotherapy.

Our approach is deliberately pragmatic. We move beyond abstract ethical discussions to outline design imperatives that correspond to real-world clinical workflows and regulatory expectations. This perspective is particularly relevant for public mental health systems, which face challenges such as long waitlists, workforce shortages, and equity mandates while needing to ensure safety, quality, and patient-centered care.

Ultimately, even when artificial intelligence systems meet the highest standards of explainability, shared decision making, electronic health record integration and harm prevention, they cannot substitute for the human element of psychotherapy (Chandler et al., 2022). The personal encounter- the ability to hold ambiguity, remain fully present, and cultivate a genuine therapeutic alliance- remains beyond the capabilities of algorithms (Sufyan et al., 2024; Frances, 2025; Poudel

et al., 2025). As technology advances, our responsibility is to ensure it serves and strengthens this relational space rather than eroding it. The future of AI-supported psychotherapy will be judged not only by its technical sophistication, but by its ability to preserve the humanity at the heart of care.

References

Aggarwal, N. K., & Bhugra, D. (2025). The ethics of artificial intelligence-based psychotherapy and the future of psychiatry. *International Review of Psychiatry, 1–*7. https://doi.org/10.1080/09540261.2025.2559108

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I.; Precise4Q consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Arvai, N., Katonai, G., & Mesko, B. (2025). Health care professionals' concerns about medical AI and psychological barriers and strategies for successful implementation: Scoping review. *Journal of Medical Internet Research, 27*, e66986. https://doi.org/10.2196/66986

Auf, H., Svedberg, P., Nygren, J., Nair, M., & Lundgren, L. E. (2025). The use of AI in mental health services to support decision-making: Scoping review. *Journal of Medical Internet Research, 27*, e63548. https://doi.org/10.2196/63548

Benrimoh, D., Whitmore, K., Richard, M., Golden, G., Perlman, K., Jalali, S., … Margolese, H. C. (2025). Artificial intelligence in depression-medication enhancement (AID-ME): A cluster randomized trial of a deep-learning-enabled clinical decision support system for personalized depression treatment selection and management. *Journal of Clinical Psychiatry, 86*(3), 24m15634. https://doi.org/10.4088/JCP.24m15634

Campellone, T. R., Flom, M., Montgomery, R. M., Bullard, L., Pirner, M. C., Pavez, A., … Darcy, A. (2025). Safety and user experience of a generative artificial intelligence digital mental health intervention: Exploratory randomized controlled trial. *Journal of Medical Internet Research, 27*, e67365. https://doi.org/10.2196/67365

Chandler, C., Foltz, P. W., & Elvevåg, B. (2022). Improving the applicability of AI for psychiatric applications through human-in-the-loop methodologies. *Schizophrenia Bulletin, 48*(5), 949–957. https://doi.org/10.1093/schbul/sbac038

Chmielowska, M., Zisman-Ilani, Y., Saunders, R., & Pilling, S. (2023). Trends, challenges, and priorities for shared decision making in mental health: The first umbrella review. *International Journal of Social Psychiatry, 69*(4), 823–840. https://doi.org/10.1177/00207640221140291

Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., & Holmes, J. H. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine, 133*, 102423. https://doi.org/10.1016/j.artmed.2022.102423

Cruz-Gonzalez, P., He, A. W., Lam, E. P., Ng, I. M. C., Li, M. W., Hou, R., … Sánchez Vidaña, D. I. (2025). Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine, 55*, e18. https://doi.org/10.1017/S0033291724003295

Dehbozorgi, R., Zangeneh, S., Khooshab, E., Nia, D. H., Hanif, H. R., Samian, P., … Lohrasebi, F. (2025). The application of artificial intelligence in the field of mental health: A systematic review. *BMC Psychiatry, 25*(1), 132. https://doi.org/10.1186/s12888-025-06483-2

Edgcomb, J. B., & Zima, B. (2019). Machine learning, natural language processing, and the electronic health record: Innovations in mental health services research. *Psychiatric Services, 70*(4), 346–349. https://doi.org/10.1176/appi.ps.201800401

El Arab, R. A., Al Moosa, O. A., & Sagbakken, M. (2025). Economic, ethical, and regulatory dimensions of artificial intelligence in healthcare: An integrative review. *Frontiers in Public Health, 13*, 1617138. https://doi.org/10.3389/fpubh.2025.1617138

European Data Protection Supervisor. (2023, November 16). *TechDispatch 2/2023: Explainable artificial intelligence (XAI).* https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf

Frances, A. (2025). Warning: AI chatbots will soon dominate psychotherapy. *The British Journal of Psychiatry, 1*–5. https://doi.org/10.1192/bjp.2025.10380

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports, 21*(11), 116. https://doi.org/10.1007/s11920-019-1094-0

Grenon, S., Hoang, K. N., Luo, S., & Koch, A. (2025). Adolescent mental health in rural settings: The role of artificial intelligence and community engagement. *Frontiers in Public Health, 13*, 1643466. https://doi.org/10.3389/fpubh.2025.1643466

Grodniewicz, J. P., & Hohol, M. (2023). Waiting for a digital therapist: Three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry, 14*, 1190084. https://doi.org/10.3389/fpsyt.2023.1190084

Grzenda, A., & Widge, A. S. (2024). Electronic health records and stratified psychiatry: Bridge to precision treatment? *Neuropsychopharmacology, 49*(1), 285–290. https://doi.org/10.1038/s41386-023-01724-y

Gual-Montolio, P., Jaén, I., Martínez-Borba, V., Castilla, D., & Suso-Ribera, C. (2022). Using artificial intelligence to enhance ongoing psychological interventions for emotional problems in real- or close to real-time: A systematic review. *International Journal of Environmental Research and Public Health, 19*(13), 7737. https://doi.org/10.3390/ijerph19137737

Gundlack J, Negash S, Thiel C, Buch C, Schildmann J, Unverzagt S, Mikolajczyk R, Frese T; PEAK consortium. (2025). Artificial intelligence in medical care—Patients' perceptions on caregiving relationships and ethics: A qualitative study. *Health Expectations*, *28*(2), e70216. https://doi.org/10.1111/hex.70216

Gutierrez, G., Stephenson, C., Eadie, J., Asadpour, K., & Alavi, N. (2024). Examining the role of AI technology in online mental healthcare: Opportunities, challenges, and implications, a mixed-methods review. *Frontiers in Psychiatry, 15*, 1356773. https://doi.org/10.3389/fpsyt.2024.1356773

Hawley, S., Yu, J., Bogetic, N., Potapova, N., Wakefield, C., Thompson, M., … Rotenberg, D. (2021). Digitization of measurement-based care pathways in mental health through REDCap and electronic health record integration: Development and usability study. *Journal of Medical Internet Research, 23*(5), e25656. https://doi.org/10.2196/25656

Humayun, A., Madawana, A. M., Hassan, A., Mahmud, A., Kamaruddin, N., Noor, S. H., … Awang Nawi, M. A. (2025). Artificial intelligence as a predictive tool for mental health status: Insights from a systematic review and meta-analysis. *PLOS ONE, 20*(9), e0332207. https://doi.org/10.1371/journal.pone.0332207

Kellogg, K. C., & Sadeh-Sharvit, S. (2022). Pragmatic AI-augmentation in mental healthcare: Key technologies, potential benefits, and real-world challenges and solutions for frontline clinicians. *Frontiers in Psychiatry, 13*, 990370. https://doi.org/10.3389/fpsyt.2022.990370

Kerz, E., Zanwar, S., Qiao, Y., & Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry, 14*, 1219479. https://doi.org/10.3389/fpsyt.2023.1219479

Khan, W. U., & Seto, E. (2023). A "Do No Harm" novel safety checklist and research approach to determine whether to launch an artificial intelligence-based medical technology: Introducing the biological-psychological, economic, and social (BPES) framework. *Journal of Medical Internet Research, 25*, e43386. https://doi.org/10.2196/43386

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the

terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Mendlovic, S. (2025). Using an AI-based system for dimensional deliberate practice in psychodynamic psychotherapy. *BRAIN: Broad Research in Artificial Intelligence and Neuroscience, 16*(3), 449–463. http://dx.doi.org/10.70594/brain/16.3/34

Mertens, L., Vandenberghe, J., Bekkering, G., Hannes, K., Delvaux, N., Van Bostraeten, P., … Vermandere, M. (2025). Navigating power imbalances and stigma in mental healthcare: Patient-reported barriers and facilitators to participation in shared decision-making in mental health care, a qualitative meta-summary. *Health Expectations, 28*(2), e70239. https://doi.org/10.1111/hex.70239

Mohsin Khan, M., Shah, N., Shaikh, N., Thabet, A., Alrabayah, T., & Belkhair, S. (2025). Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics, 195*, 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780

Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., Achtyes, E., & Bauer, M. (2022). Expectations for artificial intelligence (AI) in psychiatry. *Current Psychiatry Reports, 24*(11), 709–721. https://doi.org/10.1007/s11920-022-01378-5

Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports, 15*(1), 31652. https://doi.org/10.1038/s41598-025-17242-4

Poudel, U., Jakhar, S., Mohan, P., & Nepal, A. (2025). AI in mental health: A review of technological advancements and ethical issues in psychiatry. *Issues in Mental Health Nursing, 46*(7), 693–701. https://doi.org/10.1080/01612840.2025.2502943

Putica, A., Khanna, R., Bosl, W., & Saraf, S. (2025). Ethical decision-making for AI in mental health: The integrated ethical approach for computational psychiatry (IEACP) framework. *Psychological Medicine, 55*, e213. https://doi.org/10.1017/S0033291725101311

Slade, M. (2017). Implementing shared decision making in routine mental health care. *World Psychiatry, 16*(2), 146–153. https://doi.org/10.1002/wps.20412

Stephenson, C., Eadie, J., Holmes, C., Asadpour, K., Gutierrez, G., Kumar, A., … Alavi, N. (2025). Evaluation of an artificial intelligence and online psychotherapy initiative to improve access and efficiency in an ambulatory psychiatric setting: Évaluation d'une initiative de psychothérapie en ligne basée sur l'intelligence artificielle visant à améliorer l'accès et l'efficacité en milieu psychiatrique ambulatoire. *Canadian Journal of Psychiatry, 0*(0). https://doi.org/10.1177/07067437251355641

Sufyan, N. S., Fadhel, F. H., Alkhathami, S. S., & Mukhadi, J. Y. A. (2024). Artificial intelligence and social intelligence: Preliminary comparison study between AI models and psychologists. *Frontiers in Psychology, 15*, 1353022. https://doi.org/10.3389/fpsyg.2024.1353022

Sun, J., Lu, T., Shao, X., Han, Y., Xia, Y., Zheng, Y., … Lu, L. (2025). Practical AI application in psychiatry: Historical review and future directions. *Molecular Psychiatry, 30*(9), 4399–4408. https://doi.org/10.1038/s41380-025-03072-3

Tekin, Ş., & Delehanty, M. (2025). Beyond doomsday fears: Why we need to consider the potential harms of AI psychotherapy. *American Journal of Bioethics, 1*–11. https://doi.org/10.1080/15265161.2025.2457724