JNphi

From Wittgenstein's Language Games to LLMs: Representation, Meaning, and the Future of Psychotherapy

Hasan Belli[1], Fırat Belli[2], Hasan Gökçay[3]

Abstract

AI and large language models (LLMs) are rapidly taking center stage in clinical psychology and psychotherapy, raising fundamental questions about how machines represent language, emotion, and lived experience. LLMs exhibit linguistic fluency and contextual adaptability, but these capabilities are based on statistical relationships rather than human cognition, presenting significant limitations in therapeutic contexts. Drawing on Wittgenstein's concepts of language games and forms of life, this article examines representation problems in AI-assisted therapeutic dialogue and argues that LLMs can imitate but not genuinely internalize the relational and experiential dimensions fundamental to psychotherapy. The analysis explores how emotional simulation varies across LLMs, why simulated empathy diverges from therapeutic empathy, and how clinical safety concerns arise from limitations in contextual reasoning and emotional attunement. Based on a focused narrative review of 36 studies from 2015–2025, findings indicate that LLMs offer potential for assessment, diagnostic support, training, and psychoeducation, but remain limited in representing affect, cultural nuances, and embodied co-regulation essential to therapeutic relationships. The authors propose alternative systems based on hybrid neuro-symbolic architectures, multimodal affect systems, and interdisciplinary collaboration.

Key Words: large language models, psychotherapy, representation, simulated empathy, Wittgenstein, affective computing

DOI: 10.5281/zenodo.20267074

1. Introduction

Artificial intelligence (AI) has become one of the most influential technologies of the twenty-first century. AI is reshaping fields such as medicine, finance, education, and interpersonal communication

Corresponding author: Hasan Gökçay[3]
Address: [1]Istanbul Atlas University, School of Medicine, Department of Psychiatry, Bagcılar, Istanbul, Türkiye. [2]Freelance Linguist, Bahçelievler, Istanbul, Türkiye. [3]Council of Forensic Medicine, Bahçelievler, Istanbul, Türkiye.
e-mail ✉ hasangkcy@yahoo.com

(Russell, 2020). AI also offers significant opportunities in mental health and psychotherapy. Nevertheless, it is also perceived as a source of conceptual and ethical challenges. This rapid development renders it insufficient for clinicians, researchers, and philosophers to assess only technical capacities. These developments require these individuals to also reevaluate the deeper implications concerning how machines process and represent human language and emotions (Topol, 2019).

At the core of these debates lies the problem of representation how artificial systems encode linguistic, emotional, and experiential content. This issue connects contemporary AI research with long-standing philosophical discussions, particularly those shaped by Ludwig Wittgenstein and Alan Turing (Wittgenstein, 2009). In his later works, Wittgenstein argued that meaning does not reside in symbols themselves. The philosopher emphasized that meaning emerges through their use within "language games" embedded in shared human contexts. Turing, on the other hand, focused on the computability of symbols and the functional possibility of machine intelligence (Turing, 1950; Wittgenstein, 2009). Their intellectual approaches continue to inform contemporary debates regarding symbolic, connectionist, and hybrid approaches to AI (Marcus, 2020). The emergence of large language models (LLMs) such as Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) has marked the beginning of a significant breakthrough in this field. These developments have demonstrated both the remarkable strengths and fundamental limitations of modern AI (Bender & Koller, 2020). Large Language Models (LLMs) largely rely on distributed representations and the processing of large-scale information through machine training. In this process, natural language processing models are also used. LLMs can thereby produce linguistically coherent texts and answer complex questions. Furthermore, they can imitate empathic speech patterns (Devlin et al., 2018). Nevertheless, an important question remains: Do these outputs reflect genuine human understanding, or are they statistical results containing subtle details (Pennisi & Falzone, 2024)? This question becomes particularly critical in psychotherapeutic approaches, because psychotherapeutic encounters are not based solely on information exchange. This process also depends on the nuanced and authentic interpretation of emotions, intentions, and vulnerabilities (Cozolino, 2014). Recently, AI-supported cognitive-behavioral therapy and chatbot-based tools have been developed. These tools facilitate people's access to such services; at the same time, they hold the potential to reduce costs and address professional shortages (Torous et al., 2019). However, risks of algorithmic bias and privacy violations stand out as serious concerns. In addition to these, significant ethical and epistemic concerns come to the fore in areas such as the potential erosion of human connection, which lies at the heart of psychotherapy (Floridi et al., 2018). Therefore, evaluating AI

in clinical contexts requires much more than technical criteria. Deeper issues such as meaning, trust, and representation must also be discussed (Beg et al., 2024). This article aims to examine these points of intersection. It draws on philosophical frameworks, AI research traditions, and emerging clinical practices to explore how representation and language shape the future role of AI in psychotherapy.

2. Methodology

The authors attempted, as a methodological design, to synthesize and discuss LLMs, representational theory, and their contemporary approaches to psychotherapeutic applications. For this purpose, they conducted a literature review to construct a narrative that would form this synthesis. The authors performed a systematic search. The search was carried out in major academic databases such as PubMed, Scopus, APA PsycINFO, Google Scholar, and arXiv. As a search strategy, the authors used keywords including "large language models," "psychotherapy," "simulated empathy," "representation," Ludwig Wittgenstein, "affective computing," and "clinical natural language processing." The works of certain philosophers and theorists, including Ludwig Wittgenstein, were used to establish a conceptual and philosophical foundation. Furthermore, original texts were directly examined to ensure alignment with the discussions.

2.1 Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were defined in accordance with the research objectives. Empirical or conceptual peer-reviewed studies examining the use of LLMs in clinical or therapeutic contexts, as well as works addressing representational structures in AI or philosophy of language, were included. The authors considered only publications written in the English language. The authors also excluded technical engineering articles that lacked significant psychological, philosophical, or clinical relevance. The review first addressed the emergence of transformer-based models and subsequently explored their current applications in psychotherapy and cognitive science. During the full-text review stage, studies that focused solely on technical model optimization without any connection to psychotherapy, that did not contain substantial discussion of representation or therapeutic interaction, or that repeated conceptual content already represented in the existing literature were excluded.

2.2 Coding and Classification

A mixed coding strategy combining a-priori and in-vivo coding techniques was adopted to classify the selected studies. A-priori codes were established based on predefined conceptual domains, including representation, affective response modeling, clinical NLP, and philosophical grounding. During full-text analysis, in-vivo coding was applied to incorporate emergent categories related to therapist–AI interaction, simulated empathy, and language-philosophical constructs. Codes and sub-codes were iteratively refined to reflect recurring themes across the corpus, ensuring coherence with both theoretical frameworks and empirical findings.

2.3 PRISMA-Based Review Process

Following PRISMA guidelines, the systematic search initially identified 1.237 records from databases, with an additional 98 records identified from other sources. After removing duplicates, 1.150 records remained for screening. Titles and abstracts were reviewed for relevance, leading to the exclusion of 1.051 records that did not meet the inclusion criteria. Subsequently, 99 full-text articles were assessed in detail, and 63 articles were excluded for failing to meet the predefined criteria. As a result, 36 studies were included in the final qualitative synthesis. This PRISMA-guided process ensured transparency and methodological rigor, providing a solid foundation for analyzing how representational systems in AI intersect with philosophical perspectives and emerging therapeutic practices (Figure 1).

Many authors have been unable to fully situate the concept of AI within a single, universally accepted definition due to its interdisciplinary nature. This definition is relative and varies across many disciplines. Disciplines such as computer science, cognitive psychology, philosophy, and linguistics have addressed the issue, but these interpretations differ considerably in scope and emphasis. Russell and Norvig (2020) classified AI definitions into four broad categories: thinking humanly, acting humanly, thinking rationally, and acting rationally. These frameworks encompass different aspects of intelligence. These different aspects include domains such as internal cognitive processes, observable behavior, logical reasoning, or goal-directed action.

3.1 Rational Action (Logical Approach)

The rational action perspective defines AI as the capacity to make optimal decisions to achieve a goal. Rooted in symbolic AI and logical reasoning systems (Nilsson, 1998), this approach underlies decision-support tools, planning algorithms, and optimization models.

3.2 Thinking Like Humans (Cognitive Approach)

The cognitive approach aims to model human learning, memory, problem-solving, and language comprehension. Natural Language Processing (NLP) systems and cognitive architectures such as Adaptive Control of Thought – Rational (ACT-R) and State Operator, And Result (SOAR) exemplify this tradition, bridging AI research with psychology and neuroscience (Newell and Simon, 1973).

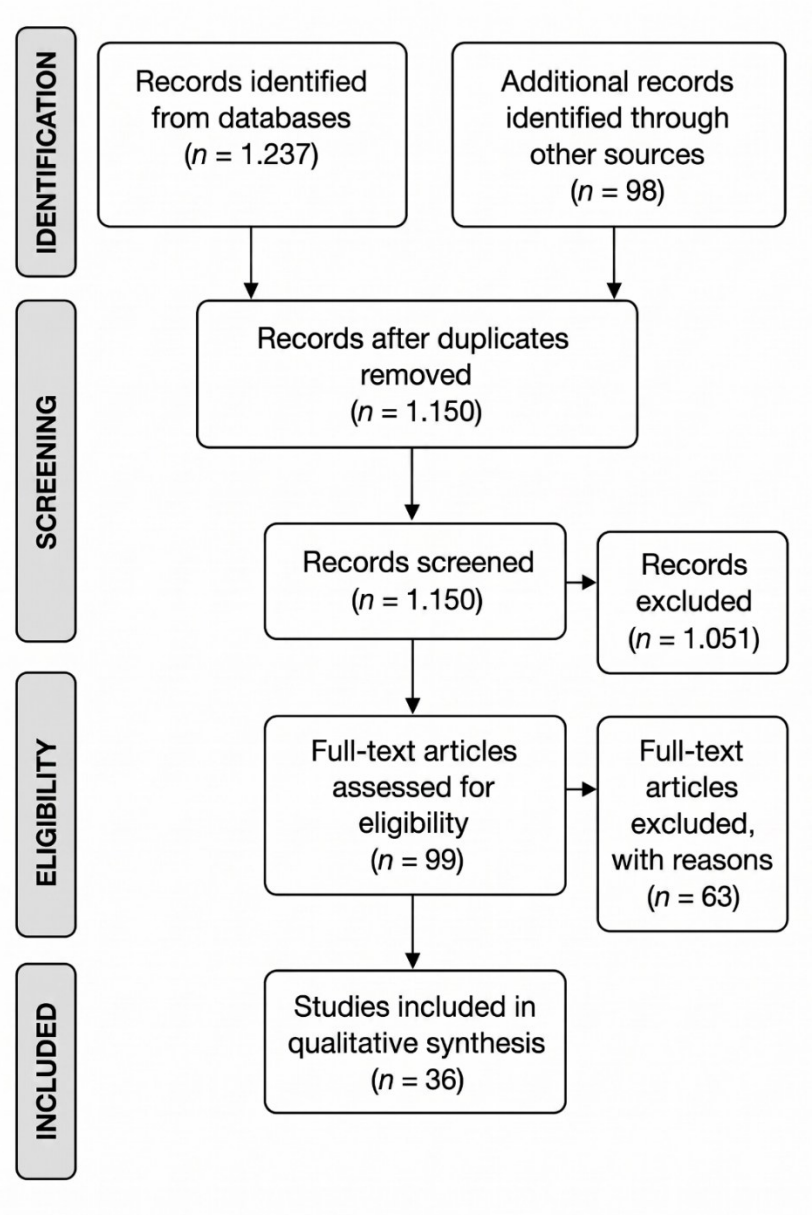

Figure 1. PRISMA Flow Diagram of the Study Selection Process

3. Definition of Artificial Intelligence

3.3 Acting Like Humans (Pragmatic Approach)

The pragmatic or behaviorist approach defines intelligence by the extent to which a machine's behavioral outputs are indistinguishable from those of a human. In this context, human-likeness is the primary criterion. Turing's famous test remains a key foundation of this view (Shieber, 2004). Modern LLMs often meet this standard in conversation-based settings. Nevertheless, the differences between imitation and genuine understanding come into question. Taken together, these perspectives reveal AI as a heterogeneous field shaped by technical, cognitive, and philosophical considerations. No single definition captures its full complexity, reflecting the ongoing tensions between symbolic and experiential models of intelligence.

4. Representation in Artificial Intelligence: Conceptual and Technical Dimensions

Representation refers to the transformation of real-world objects, events, or ideas into computationally operable structures (Brachman and Levesque, 2007). This is both a technical and philosophical challenge because representational choices determine what a system can “understand.”

4.1. Symbolic Representation

Symbolic AI encodes knowledge through explicit rules, logical structures, and semantic networks, with historical roots in Leibniz, Boole, and Frege. While such systems offer transparency and consistency, they struggle with ambiguity and uncertainty (Davis et al., 1993).

4.2. Connectionist Representation

Connectionist models conceptualize knowledge as patterns encoded in the weighted connections of artificial neural networks. Their theoretical foundations trace back to Hebbian learning principles, Rosenblatt’s early perceptron architecture, and the Parallel Distributed Processing (PDP) framework developed by Rumelhart and colleagues (Rosenblatt, 1958; Rumelhart et al., 1988). Contemporary LLMs such as BERT and GPT extend this tradition by representing linguistic and conceptual information as high-dimensional distributed vectors within deep neural architectures (Goodfellow et al., 2016).

4.3. Hybrid Representation

Hybrid architectures combine symbolic reasoning with neural flexibility, as seen in SOAR (Newell, 1994), ACT-R (Anderson, 1996), and contemporary neuro-symbolic systems. Transformer architectures (Vaswani et al., 2017) exemplify modern hybrid dynamics, enabling both contextual analysis and abstract reasoning. Hybrid approaches offer explainability alongside adaptability, making them promising for complex, context-dependent fields such as psychotherapy.

5. Artificial Intelligence and Language Models: Technical Development and Wittgenstein's Perspective

5.1 Development and Functioning of Language Models

Until the early 2010s, language models predominantly relied on statistical approaches. N-gram models calculated the probability of a word based on the preceding words and produced reasonable results for short contexts (Jurafsky and Martin, 2025). However, these patterns lack mechanisms to represent a broader context. Therefore, they struggle to maintain semantic coherence over longer sequences. With the development of neural architectures, we have witnessed a major transformation. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models improved the ability to retain information across extended sequences (Hochreiter & Schmidhuber, 1997). Yet even these architectures faced limitations such as vanishing gradients and difficulties in processing very long inputs. The breakthrough came with the introduction of the Transformer architecture in 2017. The publication of the paper titled "Attention Is All You Need" (Vaswani et al., 2017) crossed a threshold. The attention mechanism enabled models to evaluate relationships among all words in a sequence simultaneously. Furthermore, this development made it possible to generate long-range textual dependencies accurately and with high efficiency. These innovations enabled the development of tools containing billions of parameters trained on very large databases. As a result, contemporary LLMs such as GPT, BERT, Language Model for Dialogue Applications (LaMDA), and Pathways Language Model (PaLM) emerged. These models undergo two fundamental learning stages. During pre-training, they learn general linguistic patterns by predicting masked words or the continuation of a text. During fine-tuning, they adapt to domain-specific datasets, such as medical, legal, or clinical corpora, allowing them to perform highly specialized tasks. This combination of large-scale pattern extraction and domain adaptation explains their impressive performance in areas such as summarization, translation, question answering, and code generation (Devlin et al., 2018).

5.2 Wittgenstein and AI: From Language Games to ChatGPT

To understand the philosophical implications of LLMs, we need to consider them alongside approaches shaped by thinkers such as Ludwig Wittgenstein and Alan Turing. This approach situates the discussion within a broader intellectual landscape. These thinkers held various perspectives on meaning, logic, and computation. This perspective enriches contemporary interpretations of AI's potentials and limitations. Various philosophers of language have also joined this discussion. Some thinkers, even before the existence of LLMs, offered various critiques of machine activities that think and act like humans. In the 1970s, Italian philosophers of language such as Tullio De Mauro and Franco Lo Piparo argued that similar systems harbor linguistic problems. These problems include semantic limitations in texts and deficiencies in emphasizing cultural nuances. Furthermore, these thinkers emphasized that AI would struggle to deal with phenomena such as pragmatic variability (Devlin et al., 2018; Odifreddi, 2018; Wittgenstein, 1956). Their skepticism has shed light on contemporary debates. Based on these approaches, we may argue that despite the linguistic fluency of LLMs, they still lack the experiential foundations necessary for genuine understanding.

Wittgenstein's early philosophy, framed in the Tractatus Logico-Philosophicus, presented language as a system of logical pictures corresponding to reality. This picture theory resonated with early symbolic AI, which relied heavily on formal rules and logical structures. However, both approaches faltered when confronted with ambiguity, irony, metaphor, and the fluid nature of everyday speech.

In his later philosophy, Wittgenstein abandoned the picture theory and argued that meaning arises from the use of language within specific forms of life (Wittgenstein, 2009). Words function as tools in dynamic "language games," and their meaning depends on social practices rather than fixed logical structures. This idea aligns superficially with the way LLMs learn linguistic patterns from vast amounts of text: their knowledge emerges from use rather than explicit rule encoding.

However, Wittgenstein's key insight remains crucial: meaning requires participation in a shared form of life. LLMs, despite mastering statistical regularities, do not inhabit human social worlds. They manipulate signs without experiential grounding. Thus, they can simulate linguistic competence but cannot access the lived, contextual dimensions of meaning.

Turing's contributions complement this perspective. Through his functionalist view of intelligence, Turing proposed that machines need not replicate human consciousness to be intelligent; they need only perform rule-governed tasks effectively (Floyd, 2021; Murphy, 2021). This pragmatic viewpoint became foundational for modern computing.

Yet even Turing acknowledged that functional mimicry does not imply experiential understanding.

Historically, the dominance of symbolic AI ("Good Old-Fashioned AI," GOFAI) demonstrated the difficulty of encoding human knowledge through fixed rules. These systems excelled in well-structured domains but failed in open-ended, ambiguous contexts (Pennisi, 1998). Their limitations contributed to repeated "AI winters." The shift toward deep learning and Transformers marked a dramatic reversal, enabling systems capable of producing human-like text across a wide range of tasks (Pichler and Simo, 2022). Still, these achievements do not resolve the philosophical problem of grounding meaning.

Some authors have argued that LLMs such as ChatGPT do not rely on fixed semantic rules. According to this approach, LLMs perform Wittgensteinian language games by generating context-appropriate expressions. However, this similarity remains superficial. These models lack the embodied, cultural, and emotional foundation that constitutes the form of life, which gives human language its depth and richness. According to this approach, LLMs confirm Wittgenstein's insights in one respect, yet they also challenge these insights. While these models demonstrate how meaning can arise from use, they simultaneously expose the limitations of use without lived experience.

6. Large Language Models and Psychotherapy

LLMs have rapidly emerged as influential tools in psychotherapy research and practice. Their ability to process extensive conversational context and engage in multi-turn reasoning allows them to mimic the interactive and evolving nature of therapeutic dialogue more effectively than earlier computational systems (Na et al., 2025; Rasool et al., 2024).

6.1 Assessment

LLMs can extract symptom cues from patient narratives, detect suicidal ideation in online content, and identify cognitive distortions, enabling early detection and intervention (Bao et al., 2024; Cheng et al., 2024; Tu et al., 2024)

6.2 Diagnosis

In diagnostic tasks, LLMs synthesize subjective and objective information across dialogue turns. They support both static diagnoses (e.g., depression detection) and dynamic diagnostic reasoning during real-time interactions (Jiang et al., 2024; Ren et al., 2024).

6.3 Treatment Support

Today, LLMs have begun to address various needs in psychotherapeutic contexts. Moreover, they are increasingly undertaking more complex roles. These models can simulate dialogues based on Cognitive Behavioral Therapy (CBT). Furthermore, they can generate responses that resemble context-appropriate empathic capacity. They can function as a kind of virtual therapist (S. Lee et al., 2024; Nie et al., 2024). In parallel, they also have the capacity to correct therapists' suboptimal expressions. They work as clinical assistants that support treatment adherence by producing reframes aligned with CBT principles (Maddela et al., 2023; Welivita & Pu, 2023). In addition to these roles, LLMs serve as simulated patients in clinician training environments. They offer controlled and pedagogically valuable interaction scenarios for novice practitioners (Chaszczewicz et al., 2024). Furthermore, these models can automatically monitor session quality. Researchers also use them as assessment tools for longitudinal analysis of speech patterns throughout therapeutic processes (Y. Lee et al., 2024). Despite all these developments, some significant problems persist. Current LLM applications exhibit a fragmented nature. They have narrow scopes of use. For instance, they focus on depression and anxiety, and the integration of various psychotherapeutic theories remains limited (Hua et al., 2024).

Additional concerns include linguistic bias, cultural insensitivity, and insufficient diagnostic reliability all of which complicate clinical adoption (Hua et al., 2024; Lawrence et al., 2024). Future progress requires continuous multi-stage modeling, psychologically grounded adaptability, and broader inclusion of underrepresented diagnostic categories (Abdelkadir et al., 2024; Kim et al., 2025).

7. Large Language Models and the Simulation of Emotions in Psychotherapy

The simulation of emotions by LLMs appears promising. However, from a deep perspective, it remains problematic. We should note that these models differ from earlier rule-based systems. LLMs can engage in extended, context-rich dialogues that resemble therapeutic interactions (De Duro et al., 2025). Nevertheless, their emotional expressions vary significantly depending on the model architecture and training data. For example, even when asked to simulate depressive symptoms, GPT models often consistently produce a positive tone. Claude Haiku, in contrast, generates more pessimistic language aligned with depressive cognition. Such differences reveal that LLMs reflect statistical biases rather than genuine emotion-based clinical understanding (Sebastiano et al., 2024). We cannot evaluate such productions of the human mind solely with statistical data. Cognitive-network analyses have shown that LLMs expand their

lexical networks throughout a conversation. However, they do not deepen semantic or emotional complexity (Stella, 2020). In contrast, human therapists can selectively focus on clinically relevant themes such as hopelessness, rumination, or maladaptive thoughts (Neenan & Dryden, 2020). Empathy generated by LLMs tends to be general, repetitive, and insufficiently targeted.

Research also frequently emphasizes the risks in problematic areas. LLMs may fail to implement necessary interventions in cases of suicidal ideation (De Choudhury et al., 2023). Furthermore, they lack the contextual awareness required to provide appropriate cognitive reframing (De Freitas et al., 2024). The potential misalignment between linguistic empathy and therapeutic empathy is quite significant. In general, LLMs offer valuable tools for training and low-intensity interventions. However, we must exercise great caution in areas such as their inability to draw from lived emotional experience and their failure to meet clinical safety standards. Future systems must move beyond superficial emotional simulation. To achieve interactions that approach genuine therapeutic sensitivity, we must integrate insights from cognitive psychology and affective science (Bertolazzi et al., 2023; Stella et al., 2023).

8. Conclusion

The problem of representation in AI becomes particularly critical when the focus shifts to psychotherapy. In many other areas of medicine, representation primarily concerns data structures, biological markers, or diagnostic categories. By contrast, psychotherapy is fundamentally concerned with the representation of lived experience, emotions, and the complex dynamics of human relationships (Topol, 2019). The therapeutic process involves transforming inner, often unconscious, material into words, symbols, and shared meanings. Against this background, the limitations of AI in representing both knowledge and affective states become especially apparent.

Wittgenstein's philosophy of language offers a unique lens through which to see these limitations (Wittgenstein, 2009). His well-known claim that "the limits of my language mean the limits of my world" carries significant importance in this context. This statement directly addresses the weaknesses of LLMs. We should note that these models have the capacity to produce coherent and context-appropriate sentences. However, they do so by using probabilistic relationships rather than drawing from real-life experiences (Bender & Koller, 2020). In psychotherapy, silence, hesitations, bodily posture, and subtle changes in speech tone can carry more meaning than explicit speech. This reveals a fundamental gap in these models' reliance on superficial patterns. An LLM may respond to a patient's expression of sadness with a grammatically correct and ostensibly empathetic statement. Yet it lacks the embodied resonance, shared emotional depth, and past

experiences that characterize a human therapist's attunement (Pennisi & Falzone, 2024).

From a neuroscientific perspective, psychotherapy is not merely an exchange of information. Psychotherapy is a process of interpersonal co-regulation. This process involves emotional resonance between therapist and client, limbic system synchronization, autonomic nervous system attunement, and rich nonverbal communication (Cozolino, 2014). Current artificial intelligence systems based on natural language processing and statistical modeling cannot replicate this neurobiological reciprocity (Sebastiano et al., 2024). The absence of genuine empathetic regulation also raises significant ethical concerns. It remains a mystery whether patients will truly feel understood. Mechanical responses can evoke feelings of alienation and disconnection (Floridi et al., 2018).

Despite these concerns, the future of AI in psychotherapy is not without promise. Integrating multimodal data streams such as vocal prosody, facial expression, and physiological signals into AI architectures could partially enhance representational depth. Combining NLP-based outputs with affective computing and multimodal fusion may allow more accurate approximations of emotional states by capturing prosodic and embodied cues beyond the semantic level (Picard, 2000; Poria et al., 2017). Such advances would move AI systems from functioning as mere linguistic mirrors toward serving as more holistic support tools. However, even advanced multimodal representation systems remain at the center of ethical debates. The psychotherapeutic relationship is fundamentally based on trust, privacy, and a sense of human-centered security. Entrusting this relational space entirely to AI could raise serious concerns. The characteristic of psychotherapy as a product of mutual interaction could be compromised, potentially reducing the process to mere information exchange (Laçin et al., 2025; Turkle, 2015). Therefore, AI should not be viewed as a complete replacement for therapists. It should be designed as a complementary resource. In practice, this could help therapists monitor emotional developments during sessions and identify suicide risk indicators. Additionally, the process may incorporate AI systems that assist in providing guided CBT modules between sessions. However, the relational core of therapy must remain firmly human-centered (Lake et al., 2017; Torous et al., 2019).

Based on these discussions, we can make some predictions about the future. The representation problem highlights the need for more sophisticated and philosophically grounded AI architectures. Psychotherapy is a process that integrates linguistic, emotional, physical, and cultural dimensions. This process requires multidimensional representations of human existence. Current BDM models only scratch the surface of this complexity. Therefore, future AI research must combine symbolic approaches with deep learning.

This integration should aim for hybrid systems that explicitly incorporate cultural and philosophical frameworks into model design. Such a hybridization could help bridge the gap between purely statistical outputs and the nuanced, context-dependent meaning-making processes required by psychotherapy (Lake and Murphy, 2021; Marcus, 2020; Torous et al., 2019).

In summary, the problem of representation in AI presents unique challenges in both philosophical and clinical contexts. Wittgenstein's concept of language games reminds us that meaning is context-dependent. This approach suggests that meaning emerges not from isolated symbols but from shared practices (Wittgenstein, 2009). In other words, meaning cannot fully emerge without the living reflections of human relationships. The psychotherapy process exemplifies this insight very well. It is less about information transfer and more about the co-construction of meaning in the presence of another person. BDM systems can simulate certain aspects of this process, but they remain limited in capturing its essence. In psychotherapy, AI should not be a force that erodes the relational core of therapy. For it to truly become a supportive tool in this process, certain conditions must be met. These conditions are intertwined with ongoing development. They include the integration of multimodal data, the design of hybrid representational systems, and the careful preservation of human presence at the center of therapeutic work (Beg et al., 2024).

Acknowledgments

The authors thank the researchers whose valuable work contributed to this review.

Conflict of Interest

The authors declare no known competing financial interests or personal relationships that could have influenced this work.

Abbreviations

Artificial intelligence: AI

Adaptive Control of Thought – Rational: ACT-R

Bidirectional Encoder Representations from Transformers: BERT

Cognitive Behavioral Therapy: CBT

Generative Pre-trained Transformer: GPT

Good Old-Fashioned Artificial Intelligence: GOFAI

Large Language Models: LLMs

Language Model for Dialogue Applications: LaMDA

Long Short-Term Memory: LSTM

Natural Language Processing: NLP

Pathways Language Model: PaLM

Parallel Distributed Processing: PDP

Recurrent Neural Networks: RNNs

State Operator, And Result: SOAR

References

Abdelkadir NA, Zhang CC, Mayo N, Chancellor S. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. Proc North Am Chapter Assoc Comput Linguistics Hum Lang Technol. 2024;2:672-680.

Anderson JR. A Simple Theory of Complex Cognition. American Psychologist 1996;51:355–65.

Bao E, Pérez A, Parapar J. Explainable depression symptom detection in social media. Health Inf Sci Syst 2024;12:1–18.

Beg MJ, Verma M, M VCKM, Verma MK. Artificial Intelligence for Psychotherapy: A Review of the Current State and Future Directions. Indian J Psychol Med 2024.

Bender EM, Koller A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of the Annual Meeting of the Association for Computational Linguistics 2020:5185–98.

Bertolazzi L, Mazzaccara D, Merlo F, Bernardi R. ChatGPT's Information Seeking Strategy: Insights from the 20-Questions Game 2023:153–62.

Brachman R, Levesque H. Knowledge Representation and Reasoning (The Morgan Kaufmann Series in Artificial Intelligence) 2007:381.

Chaszczewicz A, Shah RS, Louie R, Arnow BA, Kraut R, Yang D. Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors. Proceedings of the Annual Meeting of the Association for Computational Linguistics 2024;1:4130–61.

Cheng Z, Cheng Z-Q, He J-Y, Sun J, Wang K, Lin Y, et al. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning 2024.

De Choudhury M, Pendse SR, Kumar N. Benefits and Harms of Large Language Models in Digital Mental Health 2023.

Cozolino L. The Neuroscience of Human Relationships: Attachment and the Developing Social Brain (Second Edition) (Norton Series on Interpersonal Neurobiology) 2014:632.

Davis R, Shrobe H, Szolovits P. What Is a Knowledge Representation? AI Mag 1993;14:17–33.

De Freitas J, Uğuralp AK, Oğuz-Uğuralp Z, Puntoni S. Chatbots and mental health: Insights into the safety of generative AI. Journal of Consumer Psychology 2024;34:481–91.

Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2018;1:4171–86.

De Duro ES, Improta R, Stella M. Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing LLMs like Haiku, LLaMAntino and ChatGPT against humans. Emerging Trends in Drugs, Addictions, and Health 2025;5.

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach (Dordr) 2018;28:689–707.

Floyd J. Wittgenstein's Philosophy of Mathematics. Wittgenstein's Philosophy of Mathematics 2021.

Goodfellow I, Courville A, Bengio Y. Deep Learning 2016:1–23.

Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput 1997;9:1735–80.

Hua Y, Liu F, Yang K, Li Z, Na H, Sheu YH, et al. Large Language Models in Mental Health Care: a Scoping Review. Curr Treat Options Psychiatry 2024;12.

Jiang Y, Shen Q, Lai S, Qi S, Zheng Q, Yao L, et al. Copiloting Diagnosis of Autism in Real Clinical Scenarios via LLMs 2024.

Jurafsky D, Martin JH. Speech and Language Processing 2025. https://web.stanford.edu/~jurafsky/slp3/ (accessed August 29, 2025).

Kim Y, Jeong H, Chen S, Li SS, Lu M, Alhamoud K, et al. Medical Hallucinations in Foundation Models and Their Impact on Healthcare 2025.

Laçin S, Belli H, Laçin S. ARTIFICIAL INTELLIGENCE IN PSYCHIATRY: APPLICATIONS AND CHALLENGES. Atlas Journal of Medicine 2025;5:224–33.

Lake BM, Murphy GL. Word Meaning in Minds and Machines. Psychol Rev 2021;130:401–31.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. Behavioral and Brain Sciences 2017;40.

Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M. The Opportunities and Risks of Large Language Models in Mental Health. JMIR Ment Health 2024;11:e59479.

Lee S, Kang J, Kim H, Chung K-M, Lee D, Yeo J. COCOA: CBT-based Conversational Counseling Agent using Memory Specialized in Cognitive Distortions and Dynamic Prompt 2024.

Lee Y, Goldwasser D, Reese LS. Towards Understanding Counseling Conversations: Domain Knowledge and Large Language Models 2024:2032–47.

Maddela M, Ung M, Xu J, Madotto A, Foran H, Boureau YL. Training Models to Generate, Recognize, and Reframe Unhelpful Thoughts. Proceedings of the Annual Meeting of the Association for Computational Linguistics 2023;1:13641–60.

Marcus G. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence 2020.

Murphy PA. When Alan Turing and Ludwig Wittgenstein Discussed the Liar Paradox | by Paul Austin Murphy | Cantor's Paradise 2021. https://www.cantorsparadise.com/when-alan-turing-and-ludwig-wittgenstein-discussed-the-liar-paradox-3c2de0ff09d1 (accessed August 29, 2025).

Na H, Hua Y, Wang Z, Shen T, Yu B, Wang L, et al. A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions 2025:7362–76.

Neenan M, Dryden W. The Downward Arrow Technique. Cogn Behav Ther 2020:184–5.

Newell A, Simon H a. Human Problem Solving: The State of the Theory in 1970. American Psychologist 1973;26:145–59.

Newell Allen. Unified theories of cognition 1994:1–576.

Nie J, Shao H (Vera), Fan Y, Shao Q, You H, Preindl M, et al. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. ACM Trans Comput Healthc 2024;1.

Odifreddi Piergiorgio. Il dio della logica : vita di Kurt Gödel. Longanesi; 2018.

Pennisi A. Psicopatologia del linguaggio: storia, analisi, filosofie della mente. Carocci; 1998.

Pennisi A, Falzone A. Wittgenstein, Turing and AI. An interview with ChatGPT. Rivista Italiana Di Filosofia Del Linguaggio 2024:187–98.

Picard RW. Affective computing. MIT Press. The MIT Press; 2000.

Pichler A, Simo S. Wittgenstein and Artificial Intelligence: Towards an update. Skjolden 2022.

Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 2017;37:98–125. https://doi.org/10.1016/J.INFFUS.2017.02.003.

Rasool A, Shahzad MI, Aslam H, Chan V, Arshad MA. Emotion-Aware Embedding Fusion in LLMs (Flan-T5, LLAMA 2, DeepSeek-R1, and ChatGPT 4) for Intelligent Response Generation 2024.

Ren C, Zhang Y, He D, Qin J. WundtGPT: Shaping Large Language Models To Be An Empathetic, Proactive Psychologist 2024.

Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65:386–408.

Rumelhart DE, McClelland JL, Group the PR. Parallel Distributed Processing, Volume 1 Explorations in the Microstructure of Cognition: Foundations. A Bradford Book 1988;1:576.

Russell SJ (Stuart J. Artificial intelligence a modern approach Stuart J. Russell and Peter Norvig. Artificial Intelligence : A Modern Approach 2020:1–1115.

Sebastiano E, Duro D, Improta R, Stella M. Network science highlights the emotional structure of counselling conversations simulated by Large Language Models and humans 2024.

Shieber SM. The Turing test : verbal behavior as the hallmark of intelligence - A Bradford book 2004.

Stella M. Text-mining forma mentis networks reconstruct public perception of the STEM gender gap in social media. PeerJ Comput Sci 2020;6.

Stella M, Hills TT, Kenett YN. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. Proc Natl Acad Sci U S A 2023;120.

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.

Torous J, Andersson G, Bertagnoli A, Christensen H, Cuijpers P, Firth J, et al. Towards a consensus around standards for smartphone apps and digital mental health. World Psychiatry 2019;18:97–8.

Tu S, Powers A, Merrill N, Fani N, Carter S, Doogan S, et al. Automating PTSD Diagnostics in Clinical Interviews: Leveraging Large Language Models for Trauma Assessments 2024.

Turing AM. COMPUTING MACHINERY AND INTELLIGENCE. Mind 1950;LIX:433–60.

Turkle S. Reclaiming conversation: The power of talk in a digital age. New York: Penguin Press; 2015.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Adv Neural Inf Process Syst 2017;30:5998–6008.

Welivita A, Pu P. Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy. Proceedings of the Annual Meeting of the Association for Computational Linguistics 2023:5411–32.

Wittgenstein L. Remarks on the Foundations of Mathematics. vol. IX. Cambridge, MA: MIT. 1956.

Wittgenstein Ludwig. Philosophical Investigations. Chichester/Malden, MA: Wiley-Blackwell; 2009.