

Emotional AI as Psychological Support: Functional Limitations and the Reshaping of Relational Expectations

Scoping Review

Liting Xiao *

College of Business, City University of Hong Kong, Kowloon Tong, Hong Kong SAR, China

* Corresponding Author Email: ltxiao2-c@my.cityu.edu.hk

Abstract. Artificial intelligence chatbots and conversational agents are being used more and more often to provide emotional support in mental health. A scoping review of interdisciplinary studies published from 2018 to 2025 was conducted to identify structural and functional deficiencies in emotional AI for mental health support. Five connected thematic deficiencies have been identified: a predictive rather than an interpretive understanding, instability of the long-term dialogic context, absence of clinical authority and accountability, affirmation-oriented interaction design, and the perceived sufficiency of AI support. These topics are then related to user experience and, considering emotion, relationship expectations, and professional help-seeking, interpreted. Although emotional AI is more readily available for people, it also has some drawbacks; thus, people's expectations of emotional support may be reshaped, and they may fail to receive timely professional help. Implications for research, clinical practice and system design are proposed.

Keywords: Emotional AI; Mental health chatbots; Conversational agents; Therapeutic relationship; Help-seeking; Scoping review.

1. Introduction

Recently, artificial intelligence in mental health care has been developing and advancing at a high speed. Mental health chatbots and conversational agents are now being used to some extent as alternatives to or supplements for traditional therapy. These technologies can provide 24/7 support at a lower cost and are free from the stigma that may prevent people from seeking professional help [1-2]. This growth occurs under the circumstances of an urgent need, as one in four adults suffers from mental illness. Mental health services are only available to 15-45% of those in need of support, and this figure is even lower in low-resource areas [1]. Given the above circumstances, AI-assisted emotional support has begun to emerge as a potential remedy for the current problems of mental health access and service capacity.

Although emotional AI has gained attention in recent years for mental health support, research on the long-term psychological and clinical consequences of its application has not kept pace with the speed of development and application. Most research on emotional AI has focused on user satisfaction, quick symptom relief and convenience. Few studies have explored the relational and therapeutic implications of long-term use of emotional AI systems.

This review has two objectives. First, it aims to identify, through a scoping review of interdisciplinary literature, key thematic limitations of emotional AI in mental health support, particularly in relation to functionality and structure. Second, it aims to interpret these identified themes within the context of user experience, with particular attention to users' emotional experiences, relational expectations, and help-seeking practices.

The reviewed literature suggests five interconnected thematic limitations: predictive rather than interpretive understanding, instability of long-term dialogic context, absence of clinical authority and accountability, affirmation-based dialog design, and the perceived sufficiency of AI support. These themes are based on research on language model limitations, conversational agents in healthcare, therapeutic relationships, clinical accountability, and chatbot effectiveness in mental health contexts [1, 3-7].

The following sections outline the scoping review methodology, present the five thematic findings, and discuss their implications for emotional support, relational expectations, and professional help-seeking.

2. Methodology

2.1. Scoping Review Methodology

This study utilized a scoping review method to map and synthesize the interdisciplinary literature on emotional AI in mental health support. The area is broad, rapidly evolving, and has many forms of evidence in artificial intelligence, clinical psychology, digital mental health, and human-computer interaction. While systematic reviews assess the effectiveness of interventions, scoping reviews may be better suited to spotting themes, concepts, and research directions that emerge from these new areas of study.

Articles in the fields of artificial intelligence, mental health and human-computer interaction have been retrieved from PubMed, PsycINFO, ACM Digital Library and IEEE Xplore. Search terms combined three concept clusters: (1) technology-related terms such as “artificial intelligence,” “chatbot,” “conversational agent,” and “language model”; (2) mental health and support-related terms such as “mental health,” “psychological support,” “therapy,” and “emotional support”; and (3) interaction-related terms such as “relationship,” “interaction,” “help-seeking,” and “limitations.” Only English-language publications from 2018 to 2025 were included in order to capture recent developments in conversational AI and emotional support technologies.

Articles were included if they addressed the capabilities, limitations, or relational implications of AI-based conversational systems used in mental health and emotional support. Both empirical and conceptual literature was included where relevant to the objectives. Studies focused only on non-interactive digital tools, clinical decision-support systems that did not directly engage users, or AI applications unrelated to emotional or psychological support were excluded from the review. Grey literature was generally avoided, save for clinical guidelines published by authoritative organizations, such as NICE.

The literature was reviewed through thematic synthesis. This study identified the common themes and conceptual issues in the literature during the review and clustered them into related thematic categories. These topics included understanding predictively rather than interpretively, instability of long-term contextual interaction, lack of clinical authority and accountability, affirmation-oriented interaction design, and the potential of emotional AI to affect users’ relationship expectations and help-seeking behaviors. The thematic synthesis also allowed examination of how the technical features of emotional AI may impact psychological and relational aspects of emotional AI.

2.2. PRISMA-ScR Flow Diagram

The literature identification and selection process is presented in Fig. 1.

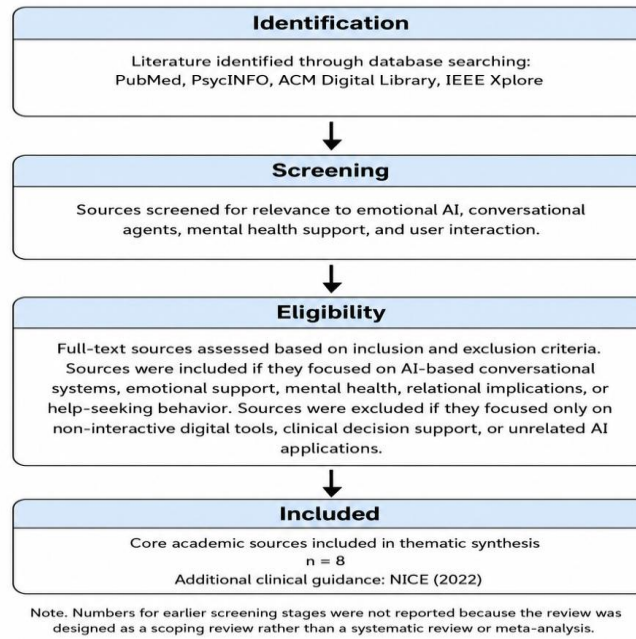


Fig. 1 PRISMA-ScR flow diagram of the literature identification and selection process.

2.3. Literature Summary Table

Table 1 summarizes the core academic sources and clinical guidance used in the review and clarifies how each source contributes to the thematic synthesis.

Table 1. Summary of core sources included in the thematic synthesis.

Source	Type / Context	Key Focus	Relevance to This Review
Abd-Alrazaq et al. (2020) [1]	Systematic review and meta-analysis	Effectiveness and safety of mental health chatbots	Supports discussion of short-term symptom outcomes, evidence limitations, and the difference between statistical and clinical significance.
Bender et al. (2021) [3]	Conceptual / AI ethics paper	Large language models as probabilistic language systems	Supports the theme of predictive rather than interpretive understanding.
D’Alfonso (2020) [2]	Review / digital mental health	AI applications in mental health care	Provides background on the growth, promise, and limitations of AI in mental health contexts.
Fiske et al. (2019) [4]	Ethical analysis	Embodied AI in psychiatry, psychology, and psychotherapy	Supports discussion of accountability, duty of care, crisis intervention, and integration with clinical systems.
Jacob et al. (2025) [8]	Digital communication research	Proattitudinal information and belief-shaping effects of language models	Supports the concern that AI interaction may reinforce user perspectives and shape expectations.
Laranjo et al. (2018) [5]	Systematic review	Conversational agents in healthcare	Supports limitations of healthcare chatbots in handling complex and free-flowing emotional interaction.
Liu et al. (2024) [6]	Empirical AI paper	Long-context retrieval and “lost in the middle” effect	Supports the theme of context instability and limits of long-term conversational coherence.
NICE (2022) [9]	Clinical guideline	Depression treatment and management	Provides clinical context for psychotherapy duration, professional treatment standards, and limits of AI substitution.
Norcross & Lambert (2019) [7]	Psychotherapy research book	Therapeutic relationship and psychotherapy outcomes	Supports the importance of sustained therapeutic alliance and relational continuity.

3. Thematic Findings

3.1. Predictive Rather than Interpretive Understanding

The first theme concerns the distinction between predictive language generation and interpretive psychological understanding. Emotional AI systems can produce contextually plausible responses based on patterns in large datasets; nonetheless, such an ability must not be confused with the ability to comprehend the meaning behind a person's emotional expressions psychologically. Such a differentiation is especially important within the realm of psychological well-being, because, more often than not, emotional distress often requires interpretation of intention, context, risk, and relational meaning, rather than the generation of linguistically appropriate responses alone [3, 5].

Emotional AI appears to provide support. However, its functionality is based on probabilistic modeling of linguistic forms rather than an accurate interpretation of their psychological meaning. According to Bender et al. [3], large language models are best described as "*stochastic parrots*", a term that warrants careful examination. The word *stochastic* refers to the fact that model outputs are generated through probabilistic sampling: the system draws from learned probability distributions over token sequences, producing outputs that reflect past patterns in training data but are not determined by any underlying comprehension. The word parrot captures the imitative quality of this process, just as a parrot that has learned to say "Polly wants a cracker." reproduces the sounds based on previous patterns but does not understand the meaning or hunger of "Polly wants a cracker" or the concept of a cracker, an LLM generates emotionally resonant strings when it hears words without any understanding of distress, comfort, or the human experience they describe. When an AI gives a comforting response, it does not feel empathy. It is only performing symbol manipulation and lacks genuine "*communicative intent*."

The clinical relevance of this shortcoming can easily be seen once one investigates what the mechanisms of emotional expression actually look like in real-life human communication. For instance, imagine a user talking to the AI, after a difficult exam and saying, "I'm so angry I could explode." Linguistic context (i.e., the expression of anger and its degree) will be considered by the AI when deciding how to respond to the message. In a clinical scenario, a skilled professional will almost immediately switch the lens and ask several questions: Is this anger expression actually a manifestation of anxiety? Has the examination served as a trigger for even deeper issues relating to self-worth in the academic context? Is the level of anger being expressed appropriate, or does it seem exaggerated or unsettling? Such forms of contextual and psychological interpretation remain difficult for current AI systems to reliably perform as they require interpreting the speaker's experience of reality and are not related to simple token analysis. An additional example that will demonstrate the difference between clinical and predictive approach to emotional expression is the situation when users use extremely negative wording, like "wanting to die," in a metaphorical way in order to express frustration or exhaustion, instead of being serious about their intentions. AI systems may interpret such statements as indicators of suicidal ideation due to safety-oriented response protocols.

Such technology reduces the problem of complex psychology to statistical correlations of language use [3]. AI cannot experience life and learn about it through practice. In this case, any kind of "understanding" is merely a device created by the user. The sense of "understanding" is produced by projection; thus, people derive meaning from the text independently of the AI's objectives. Laranjo et al. [5] have found that chatbots cannot grasp the nuances of people's feelings in complex or free-form language, and their responses are often pre-scripted without any professional counselling experience. The difference between the interpretation and the prediction is not due to a lack of more training data; instead, it reflects the gap between pattern recognition and understanding. Although emotional AI can have a continuous and supportive conversation, it struggles to understand subtle changes in emotion, and thus may not be suitable for dealing with complex psychological problems or high-risk mental health situations [2, 3, 5].

3.2. Context Instability in Long-Term Dialogue

The second theme is the instability of context in extended conversation. Based on the above research, emotional AI systems are unlikely to maintain the kind of context-awareness needed for extended psychological support. This is a serious problem because psychotherapy is not a one-time occurrence; rather, it is a continuous process in which the therapist gradually comes to know the patient's mind [7]. According to research on the actual effects of psychotherapy, a good therapeutic relationship that has been built over a long period of time can help people improve [7]. A therapist who has worked with a patient for over 12 sessions can observe changes in the patient's mood and deviation from the previous situation, as well as continue the narrative built by the patient.

The first source of instability is partiality in the data that AI systems receive. Users generally use emotional AI when they are in a bad mood, such as after an argument, at night without sleep, or when they are anxious. These are the times when people's emotions are most intense; they are not in their normal state. An AI system that records these interactions will create an unbalanced user profile that focuses on high-stress periods. When the same user returns to the system after the crisis has passed, they are more relaxed and have solved the problem; at this time, current AI systems may be unable to continuously incorporate such contextual changes into the ongoing representation of the user's psychological state. At that time, the emotional AI system may only respond to the current input and may not have the continuity of interpretation that a human therapist provides. Furthermore, AI systems are not aware of the circumstances in which users are communicating; for instance, the AI cannot know whether a person is intoxicated, drowsy or emotionally unstable at the time of use. Based on the tone of voice, physical characteristics and previous knowledge of the client, the clinician will conduct the above observation. Most of the text-based interaction environments for AI systems are limited to linguistic input, and they do not have access to extensive context or other non-verbal cues available in a clinical setting.

The second type of structural instability for large language models is their technical architecture. Large language models process previous conversations in a finite context window; that is to say, only a certain amount of the preceding text can be included in the response being generated now. Based on empirical studies, this deficiency is real, not hypothetical. Liu et al. [6] have shown that language models have a U-shaped performance curve for processing long contexts; that is, their performance is highest when the relevant information is at the very beginning (primacy bias) or the end (recency bias) of the input, but it drops significantly—by more than 20% in some cases—when the model needs to retrieve information from the middle of an extended conversation. Over the course of many turns in a conversation, the beginning statements may no longer be available due to the model's limited context window. With many turns, the information that was disclosed earlier may have become less accessible or conspicuous in the following responses. Therefore, there is a problem for emotional support, as early disclosures may be clinically relevant to interpreting later emotional expressions. This is far from the standard of care for human psychotherapy; most authorities recommend a minimum of 12 sessions over the course of a year for common disorders such as depression and anxiety [9], and treatment for complex disorders may require many more sessions. As Liu and others [6] have reported, even models with extended context windows of 16K tokens may show a degradation in performance when needed to integrate information disclosed early in the treatment with the current conversational context. As the interaction lengthens, the AI's capacity for continuous and integrated support may decline; thus, this may not be suitable for prolonged psychological care.

In short, the reviewed literature raises problems with the current emotional AI system's ability to maintain continuity and context in extended psychotherapy. In other words, emotional AI can respond appropriately in the short term to a particular conversation, but it is unlikely to form a stable sense of what the user is feeling over an extended period. With extended and more complicated interactions, this deficiency will likely be more pronounced [6, 7, 9].

3.3. Absence of Clinical Authority and Accountability

The third theme is the absence of clinical authority and professional responsibility for emotional AI systems. Emotional AI systems do not fall under the same legal, institutional and professional framework as human therapists for clinical mental health treatment. Although human therapists are subject to a number of requirements, such as licences, ethical codes and responsibilities in case of a crisis, emotional AI systems do not fall under these frameworks at present [2, 4].

For those who are immediately at risk of serious psychological harm, the problem cannot be solved simply by enhancing the algorithm. Fiske and colleagues [4] have shown that mental health professionals are part of a system for duty of care, referral, supervision and emergency response. Emotional AI systems may be able to detect signs of distress or offer crisis support, but they generally lack the power to take independent protective measures or organise care through formal clinical pathways. Therefore, the authors have posed a serious question: when a therapy bot detects a serious risk, who is held accountable, and how should referral and duty of care be managed?

Another problem is how to ensure the accountability of emotional AI systems. Human mental health practitioners are licensed and regulated by professional organisations, and are also subject to disciplinary action if they fail to meet the standards of professionalism; thus, a system of professional responsibility has been established. Accountability should specify who will be held responsible for failing to meet the professional standard or causing harm. As shown by Fiske and others [4], it is still unclear how to apply AI's duty of care and any code of conduct for the reporting of harm. If an emotional AI system fails to meet a user's expectations or provides incorrect assistance, it is unclear who will bear the legal responsibility for such an event: the developer, the company, the platform owner or the user?

None of the above problems can be solved simply by increasing the technical capabilities mentioned earlier. Although it may be possible to develop a future version of AI that can reach a level of comprehension and coherence comparable to that of humans, such systems would not be feasible for clinical application without fundamental changes to the law, professional regulations and healthcare infrastructure. In short, based on the above studies, emotional AI is not yet subject to the same legal, professional and institutional oversight as traditional mental health services.

3.4. Affirmation-Oriented Interaction Design

The fourth is an affirmation-oriented interaction design. Scholars in this field have been studying whether the goal of design for emotional artificial intelligence (AI) is to provide support and comfort for users, and whether such a design intention alters users' expectations of emotional support [4, 8]. The design goals of emotional AI systems may prioritize user satisfaction, engagement, retention and positive emotions. The above design orientation leads to what can be called affirmation bias, and this term needs closer examination. "Affirmation" refers to the situation where a user's emotions and opinions are affirmed and recognised, and the response is accepted without dispute. "Bias" is a systematic type of this tendency; it is not random but is created by the goals of the design at the system level. Therefore, affirmation bias is an organised tendency to favour validation and reassurance over responses that may challenge, complicate or clinically reframe the user's self-presentation [4, 8].

This design orientation matters because emotional support is not always the same as immediate emotional comfort. Clinicians may occasionally create a therapeutic tension through questioning, boundary-setting or subtle reframing of a client's self-narrative in human therapy. These types of interaction may be uncomfortable, but they may also be necessary for clinical reflection and learning [7]. Emotional AI systems may be unable to distinguish between instant emotional comfort and other kinds of support that are more helpful for long-term mental health.

Emotional AI's design trend of affirmation is not without reason. It is the result of making users happy and retaining them. Therefore, emotional AI may create a space that values validation over challenge, and the emotional comfort of affirmation is given more weight than critical reflection. Too many positive interactions may reshape the users' expectations of emotional support. If every

expression of distress is validated, users may come to perceive challenge, including therapeutic challenge, as a deficiency in support; thus, they will be less willing to engage in deeper relationships or clinical interactions that involve discomfort, introspection and modification [7,8].

3.5. Perceived Sufficiency of AI Support and Help-Seeking Implications

The fifth theme concerns the perceived sufficiency of AI support and how it affects the users' propensity to seek professional help. The reviewed literature is concerned about the potential effects that emotional AI may have on users' perceptions concerning whether their emotional needs were adequately met through immediate, validating and responsive support [1, 4].

Another frequently mentioned problem in the selected papers is that emotional AI may seem to offer sufficient emotional support for people in distress. Sufficient in this context should be understood not from the perspective of clinical practice, but from the users' own feelings. Here, "perception" refers to the user's subjective sense that their distress has been alleviated; it is not a clinical determination that the root cause of the problem has been solved. This feeling may be prompted more readily by AI's affirmation-based interactions. As shown in Section 3.4, emotional AI may only offer temporary emotional comfort and a sense of being "cared for", but fail to address the root causes of problems.

This sense of adequacy may delay seeking professional help. AI can be used to address the moderate-symptom group as a first step. Its empathetic language may be too soothing and thus prevent a person from seeking medical help. For users in a minor but temporary crisis, this is not a concern. However, for users with symptoms in the early stages of diseases that require professional help, such as persistent depression or newly appearing anxiety disorders, AI's reassurance may cause a "critical delay" [4] in seeking help, and the best time for early intervention may have passed.

The above problems have supporting evidence. Abd-Alrazaq and colleagues [1] conducted a systematic review of chatbots for mental health and reported that although some studies showed statistical improvements, the quality of the evidence was "very low to low" with a high risk of bias. Importantly, the observed effects often fell below the threshold of "clinical significance", and this is a problem for chatbots as substitutes for professional care. A statistical improvement does not always equate to clinical recovery; therefore, some users who have experienced only a slight improvement may be unaware of this difference and thus believe that professional help is not necessary or that their intervention falls short of the standard of care.

Another problem is that the expectations formed by AI interaction may make it more difficult for some users to engage with a therapist. As shown in Section 3.4, due to frequent exposure to positive and low-resistance interactions with AI, users may be unprepared for therapeutic friction. During the treatment stage, they may find that the difficulties therapists present to their self-narrative or boundary-setting are disorienting, and they might mistakenly believe that clinical strictness stems from a lack of empathy due to constant affirmation by the AI. Mismatched expectations may lead to early discontinuation of therapy or disengagement from professional services.

As a result, the accessibility of emotional AI brings its own complex risks. A user may postpone seeking help because AI provides subjectively satisfying validation that the user may mistake for clinically sufficient support. When the user does seek professional help, the relationship expectations created by AI interaction may hinder the user's ability to engage in a therapeutic alliance that asks more of them. The concern is not only that AI provides support, but that this form of support may sometimes delay or complicate access to the more rigorous and professional care that users ultimately need [1, 4, 7].

4. Discussion

This scoping review identifies five interconnected aspects of emotional AI in mental health support: (1) predictive versus interpretive comprehension, (2) unstable long-term contextual dialogue, (3) the lack of clinical responsibility and accountability, (4) affirming-oriented interface, and (5) perceived

sufficiency of AI support. These aspects were backed by studies discussing the limitations of the language models used, long-context instability, clinical accountability, effectiveness of chatbots, and therapeutic relations [1, 3, 4, 6, 7].

The more general implication drawn from this analysis is the conflict between accessibility and therapeutic complexity. In this regard, emotional AI can be viewed as a tool that increases the accessibility of help through continual availability, reduction of stigma, and emotional engagement [1, 2]. However, based on the findings of the literature review, these systems differ from traditional therapeutic relationships in several important ways.

First, emotional AI systems, unlike therapists, do not have interpretive psychological understanding. Second, emotional AI systems do not establish constant longitudinal context awareness. Third, emotional AI systems operate outside the framework of accountability and intervention during a mental health emergency [3, 4, 7].

The literature review also raises the question of the relational impact of affirmation-based interface design. Emotionally responsive AI systems may be designed to maximize user interaction and provide positive experiences, which may encourage interaction through validation and reassurance [4, 8]. While this may provide immediate emotional comfort, the overuse of highly affirmative and low-resistance communication may influence users' expectations for emotional support or therapeutic interaction over time. Longitudinal evidence on these relational effects remains limited, and further studies are needed to examine how repeated interaction with emotional AI may shape users' expectations of support over time [7, 8].

Another issue raised by the research is the perception of sufficiency and help-seeking behaviors. The use of AI-mediated emotional support may provide users with a subjective sense that their needs have been met, even though their underlying difficulties may remain unresolved. The literature suggests that this perception may contribute to delays in seeking professional intervention, especially when the user experiences temporary emotional relief from the AI [1, 4]. Several studies also suggest that expectations elicited through interaction with AI may not reflect the challenges, boundaries, and relational difficulties that are encountered in professional counseling [7, 8].

The results of this review provide several implications for future research and practice. For researchers, future studies should examine how repeated interaction with emotional AI may shape users' relational expectations, emotional resilience, and professional help-seeking over time. For mental health practitioners, knowing how clients have been exposed to emotional AI could play a key role in understanding what those clients expect from emotional validation, communication, and treatment. For developers and policy-makers, the results suggest that designing applications with the goal of involving the user should also consider psychological well-being, transparency, and clinical responsibility [2, 4].

Overall, this review suggests that discussion of emotional AI should extend beyond short-term symptom reduction and user satisfaction. As emotionally responsive AI systems become an increasingly frequent part of everyday psychological support environments, future research should consider the relational, clinical, and social implications of emotionally responsive AI systems [1, 2, 4].

5. Limitations

There are several limitations to this review. First, the field of emotional AI in mental health is new and rapidly evolving, making it difficult to identify long-term empirical evidence on the psychological and relational effects of prolonged interaction with emotional AI. Much of the reviewed literature is conceptual, theoretical, or exploratory in nature, while longitudinal and large-scale clinical studies remain limited. Secondly, because of the broad scope of this review, only English-language studies published between 2018 and 2025 were included in the review; works in other languages or outside the specified period could thus have been overlooked. Thirdly, because this topic spans several fields, there is some discrepancy with regard to the terminology used and to

the focus of research. Finally, while this paper does not aim at serving as an empirical study or meta-analysis on the effectiveness of the interventions mentioned, its focus is rather on conceptual issues and new challenges arising from the literature. Therefore, some of the relational and long-term psychological implications presented in the review are theoretical and require further empirical evidence to confirm.

Research must pay closer attention to the possible influence of interaction with emotional AI on people's relationship expectations, emotional resilience, and engagement with professional help in the long run. Besides, more research is necessary to understand the differences between various types of emotional AI, interaction approaches, and clinical supervision in terms of their effect on users' well-being.

6. Conclusion

There were two reasons for the study. The first was to identify five themes of limitations in emotional AI for mental health support: (1) predictive rather than interpretive understanding, (2) instability of context in long-term dialogue, (3) absence of clinical authority and accountability, (4) affirmation-centric interaction design, and (5) perceived sufficiency of AI support. The second was to interpret the above results based on users' actual experience, focusing on their expectations concerning emotions, relationships, and needs for help.

In short, although emotionally responsive artificial intelligence for mental health support can provide prompt and empathetic assistance, it may also lead to psychological and clinical problems, as well as relational issues. Emotional AI can help people to a certain extent in the short term and extend access to care and other benefits, but it has not yet been determined whether it can replace professional medical treatment for serious illnesses.

Finally, the application of emotionally responsive AI in mental health care should not focus solely on alleviating symptoms and satisfying users. Given that emotionally responsive AI will be widely used in mental health support in the future, researchers are encouraged to employ a long-term, multi-disciplinary approach to explore how multiple interactions with emotional AI may affect users' sense of emotional support, expectations for relationships, emotional resilience and participation in professional mental health services over time. Efforts should also be strengthened to determine how interaction design, affirmation-based communication and institutional supervision shape the role of emotional AI in the whole mental health care system.

References

- [1] Abd-Alrazaq A A, Rababeh A, Alajlani M, et al. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of Medical Internet Research*, 2020, 22(7): e16021.
- [2] D'Alfonso S. AI in mental health. *Current Opinion in Psychology*, 2020, 36: 112-117.
- [3] Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021: 610-623.
- [4] Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, 2019, 21(5): e13216.
- [5] Laranjo L, Dunn A G, Tong H L, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 2018, 25(9): 1248-1258.
- [6] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. *Transactions of the Association for Computational Linguistics*, 2024, 12: 157-173.
- [7] Norcross J C, Lambert M J, eds. *Psychotherapy relationships that work: Volume 1: Evidence-based therapist contributions*. 3rd ed. New York: Oxford University Press, 2019.

- [8] Jacob C, Kerrigan P, Bastos M T. The chat-chamber effect: trusting the AI hallucination. *Big Data & Society*, 2025, 12(1): 20539517241306345.
- [9] National Institute for Health and Care Excellence. Depression in adults: treatment and management (NICE guideline NG222). 2022. <https://www.nice.org.uk/guidance/ng222>