Review

AI Chatbots in Mental Health Care: Integrative Review of Challenges and Solutions

Luke Balcombe *

Posted Date: 23 September 2025

doi: 10.20944/preprints202509.1893.v1

Keywords: mental health; suicide prevention; emotionally intelligent; AI chatbots; AI companions; AI agents; challenges; solutions

 doi:10.20944/preprints202509.1893.v1

Review

AI Chatbots in Mental Health Care: Integrative Review of Challenges and Solutions

Luke Balcombe

Australian Institute for Suicide Research and Prevention, Griffith University; lukebalcombe@gmail.com

Abstract

AI chatbots are increasingly integrated into mental health care, expanding access to support yet bringing significant ethical, clinical, and design challenges. This integrative review synthesizes empirical studies, reviews, case reports, and media articles from 2015 to September 2025, critically examining the use of both rule-based and large language model (LLM) chatbots. While AI systems show promise for screening, prevention, assessment, treatment, and emotional support, concerns remain about low user retention, privacy risks, algorithmic bias, and the provision of trauma-informed, culturally sensitive care. Phenomena such as "AI psychosis" and emotional dependency further highlight the need for robust risk assessment and regulation. The review underscores the urgency of implementing the Safe Integration of LLMs in Mental Health Care Framework as well as involving vulnerable groups in the co-design process of emotionally intelligent AI chatbots to improve accessibility, safety, and effectiveness.

Keywords: mental health; suicide prevention; emotionally intelligent; AI chatbots; AI companions; AI agents; challenges; solutions

1. Introduction

The digital transformation of mental healthcare—accelerated by the COVID-19 pandemic, shifting social dynamics, and the ubiquity of smartphones—has fundamentally altered the landscape of support for youth and other vulnerable populations [1–4]. While the accessibility and reach of digital mental health tools have expanded dramatically, challenges persist [5–8]. User retention remains perilously low, ethical and regulatory frameworks lag technological development, and AI-enabled platforms, particularly those employing Large Language Models (LLMs), struggle to deliver usable, emotionally intelligent, trauma-informed care [9–13]. LLMs are advanced generative AI programs (e.g., GPT-4) that can create text, remember context in conversations, and handle tasks like giving advice or answering questions, often used for counselling-style chats [14].

AI chatbots for mental health typically operate as rule-based or scripted dialogue systems, providing users with structured psychoeducational resources or cognitive behavioural therapy (CBT) prompts [14]. These systems do not integrate statistical algorithms or machine learning for dynamic adaptation; instead, they rely on predefined responses and always require user interaction. AI chatbots autonomously support mental health care by assisting with screening, prevention, monitoring, clinical assessment and treatment, emotional support, and companionship. AI chatbots are increasingly deployed as "socially-interactive agents," "assistants," "therapists," and "companions," spanning platforms from mobile apps and web portals to social robots [15–18].

The first randomized controlled trial of a generative AI therapy chatbot (Therabot) showed moderate symptom improvement [19]. However, public reactions were mixed—many praised its accessibility and affordability, while others raised concerns about effectiveness, ethics, and safety [20]. Overall, the comments reflected deep frustration with the current mental health system and cautious interest in AI as a potential supplement to human care.

LLM-based chatbots and hybrid systems can expand access to care, support monitoring, and offer personalized interventions [14,17]. However, they also raise concerns including algorithmic

 doi:10.20944/preprints202509.1893.v1

bias, privacy risks, and integration challenges. Cases show AI gave a false impression of consciousness—while there is no evidence that AI is conscious, "Seemingly Conscious AI" is a phenomenon [21]. This calls for ethical design, transparent development, and human oversight. Despite increasing use in mental health care (e.g. with the elderly), AI chatbots continue to face challenges arising from outstanding concerns regarding user safety, effectiveness, the provision of meaningful support, enhanced safety mechanisms, human-like memory capacities, and the ability to guide therapeutic processes [22–24].

Narrative/systematic reviews and meta-analyses on generative AI highlight the need to better understand how AI chatbots impact mental health, assess their long-term effects, and integrate LLMs responsibly within ethical frameworks [25–28]. Despite their promise, LLMs face challenges such as limited data, unreliable content, and a lack of robust safeguards, making them useful tools but not replacements for professional care. The Australian Government eSafety Commissioner's position statement [29] on generative AI highlights the urgent need for Safety by Design across the AI lifecycle—especially in sensitive domains like mental health—where risks such as emotional manipulation, inappropriate responses, and epistemic harm demand ethical, clinical, and regulatory frameworks that prioritize inclusion, transparency, and protection for vulnerable users.

Research Question: What are the most urgent ethical, clinical, and design challenges confronting AI chatbots used for mental health care and assistance? To address this, the review is structured around two sub-research questions: How do phenomena such as "AI psychosis" and emotional dependency inform risk assessment, design, and regulation? What frameworks can ensure AI chatbots are not only accessible but also safe, inclusive, and effective for the most vulnerable populations?

2. Methods

An integrative review narratively synthesized empirical studies, systematic reviews, clinical case reports, case studies and key grey literature from 2015 through September 2025. Database searches included PubMed, Scopus, PsycINFO, Web of Science, and Google Scholar. Search terms encompassed "AI chatbot mental health" OR "AI companion mental health" OR "AI therapist" OR "virtual companion mental health". All abstracts were independently assessed against the inclusion and exclusion criteria according to the 5-step amendment (see Table 1) of a modified integrative review framework [3,30]. The methodology in Table 1 was applied to critically evaluate and synthesize the reported outcomes of theoretical and empirical literature on "AI Chatbots in Mental Health Care".

Table 1. Five step integrative review literature search method.

(1)	Problem/s identification
(2)	Literature search
•	Participant characteristics
•	Reported outcomes
•	Empirical or theoretical approach
(3)	Author views
•	Clinical effectiveness
•	User impact (feasibility/acceptability)
•	Social and cultural impact
•	Readiness for clinical or digital solutions adoption
•	Critical appraisal and evaluation
(4)	Determine rigor and contribution to data analysis
(5)	Synthesis of important foundations/conclusions into an integrated summation

Inclusion criteria comprised:

 doi:10.20944/preprints202509.1893.v1

- Peer-reviewed articles and systematic reviews reporting on efficacy, safety, clinical outcomes, ethics, or real-world performance of AI chatbots and companions;
- Conference proceedings and preprints addressing technical, regulatory, or operational considerations; and
- Major news media and investigative journalism documenting real-world harms and regulatory responses.

Exclusion criteria comprised:

- Non-peer-reviewed opinion pieces, editorials, or commentaries lacking empirical data;
- Marketing materials, product advertisements, and promotional literature;
- Studies focused exclusively on non-digital interventions;
- Reports not published in English or lacking full text access;
- Duplicative analyses or secondary reviews without novel synthesis; and
- Publications with insufficient methodological detail or lacking outcome data relevant to AI chatbots or companions.

Studies were screened for relevance and synthesized thematically, with the results presented in key domains of included articles, in addition to a summary of empirical findings on AI chatbots used in mental health care (see Table 1).

3. Results

3.1. Clinical Risks, Opportunities, and Ethical Issues

Studies with conversational AI chatbots such as Wysa, Woebot and Youper established promising results in facilitating early detection, supporting engagement, and effectively delivering tailored interventions, particularly for mild-to-moderate common mental health disorders and youth cohorts [31–33].

Recent advances in deep learning have enhanced conversational fluency, context tracking, and multimodal emotion recognition [34]. Nonetheless, critical deficiencies remain:

- Transparency: Most commercial AI mental health tools are proprietary, hindering scrutiny of algorithmic bias, safety logic, and escalation protocols [35,36].
- Evaluation Gaps: Few platforms have undergone rigorous clinical evaluation, especially for high-risk or marginalized groups [11].
- Stakeholder Engagement: Co-design with lived experience is rare, perpetuating cultural mismatches and failure to recognize nuanced distress cues [37].
- Privacy and Data Security: Concerns persist regarding data use, consent, and the potential for breaches or misuse [38].

Despite these challenges, AI companions have demonstrated promise in reducing loneliness and improving self-esteem, particularly among autistic adolescents, trauma-affected individuals and older adults [398–46]. However, the absence of trauma-informed protocols and effective safeguards for vulnerable users undermines both safety and inclusivity.

There is ongoing discussion regarding the development of hybrid human-AI systems that use user-centered and culturally adapted designs to increase trust and long-term engagement [25]. Ethical considerations, cultural adaptation, and the current limitations of AI in mimicking human empathy are recognized as barriers [47].

3.2. Spectrum of AI Chatbot Applications: Strengths and Weaknesses

AI chatbot applications can be classified into three principal categories:

- Therapist chatbots (e.g., Woebot, Wysa, Youper, Ash, Therabot): Deliver accessible, personalized, structured interventions and support—often based on cognitive behavioral therapy (CBT) for treating depression and anxiety—using mood tracking, psychoeducation, and

 doi:10.20944/preprints202509.1893.v1

goal setting [19]. These tools are helpful for mild to moderate symptoms and suicide prevention, however, they face issues with semantics, bias, privacy, user experience, study design/independent evaluation and measuring the therapeutic relationship [16,24,48–52].

- Companion chatbots (e.g., ChatGPT, Replika, Character.AI): Focus on relational, emotionally attuned dialogue to reduce loneliness, foster belonging, and provide a "nonjudgmental" presence. However, they often fail to prevent algorithm bias, reinforce dependency, lack depth of understanding, can inadvertently validate maladaptive beliefs, and lack adaptability to crisis escalation and trauma [53–55]. Emotionally intelligent chatbots (e.g., Hume, Voicely, Pi) are a novel class of AI that provide empathetic and supportive interactions.
- AI Agents e.g., Self-clone Chatbots, Mental Health Task Assistants, Humanoid/Social Robots:

Self-clone Chatbots are AI agents modeled on users' own conversational and support styles—as a novel alternative to traditional therapy, designed to externalize inner dialogue and enhance emotional and cognitive engagement [56].

Mental Health Task Assistants like Mia Health [57] combine psychoeducation, journaling, and real-time analytics to support care professionals across assessment, care planning, and emotion regulation. By integrating psychological expertise with advanced AI, these systems scale efficient, responsive mental health services tailored to individual needs.

Humanoid/Social robots (e.g., Qhali/Yonbo) are interactive, embodied machines with human-like appearance and/or robot features designed to engage with humans through socially intelligent behaviors—such as speech, gestures, and emotional responsiveness—with the goal of supporting mental health and well-being through companionship, motivation, and therapeutic interventions [58–62].

Generative AI-based conversational agents like ChatGPT and Replika, which autonomously generate responses using machine learning, demonstrated significantly greater reductions in psychological distress than retrieval-based agents such as Woebot and Wysa, highlighting the superior therapeutic potential of generative AI models in clinical and subclinical mental health contexts [25]. However, there is a need to better understand the underlying methods of their effectiveness, assess long term effects across various mental health and suicide outcomes, and evaluate the safe integration of LLMs in mental health care.

Large Language Model (LLM)-based chatbots, exemplified by ChatGPT-4 have increased baseline conversational "empathy" [63]. However, these LLMs remain vulnerable to:

- Hallucination and scripting errors [45];
- Loss of narrative context and memory [26];
- Bias, cultural misrecognition, and unreliable safety protocols [54,55]; and
- Failures of escalation in crisis, including lack of safeguards in cases of suicidal ideation and/or attempts [64].

Real-world use cases and case studies further uncover a range of unintended consequences, from emotional dependency and digital grief to exacerbation of psychosis and suicidal ideation—especially in vulnerable users or in the absence of robust human oversight [65,66].

3.3. AI Chatbot Methodological and Ethical Guardrails

The proliferation of mental health AI has outpaced the development of ethical, methodological, and regulatory frameworks. Safeguards emerging from the literature and expert consensus include:

- Rigorous screening tools and evidence synthesis methodologies (e.g., Mixed Methods Appraisal Tool, Joanna Briggs Institute Critical Appraisal Tool);
- Algorithmic transparency, privacy-by-design, and clear consent protocols (General Data Protection Regulation in the European Union; California Consumer Privacy Act, Health Insurance Portability and Accountability Act compliance in the US);

 doi:10.20944/preprints202509.1893.v1

- The OECD's Governing with Artificial Intelligence report outlines a comprehensive framework for trustworthy AI in government, emphasizing the importance of enablers, guardrails, and stakeholder engagement to ensure responsible and inclusive adoption [67]; and
- Standardized approaches to risk management, including human-in-the-loop systems, traceable audit trails for escalation, and continuous feedback loops [35,68].

3.4. AI Chatbot Phenomena

A growing body of investigative journalism and case studies has brought to light the darker side of AI chatbots and their impact on mental health:

- "AI psychosis":

"AI psychosis" refers to psychotic symptoms triggered or exacerbated by AI chatbot interactions—hallucinations, delusions, or a blurred sense of reality, often involving beliefs that AI is communicating directly or controlling thoughts [66,69]. Users may perceive AI as communicating secret messages, influencing their actions, or even conferring cosmic missions [70,71].

"AI psychosis" could be misinterpreted because obsessive chatbot use may trigger delusional thinking and psychotic symptoms through prolonged and emotionally immersive interactions with AI chatbots. However, it lacks the clinical features of true psychosis, which calls for more nuanced understanding and therapeutic AI design [72].

Multiple case reports describe users, often with pre-existing vulnerabilities, developing delusional beliefs or psychotic episodes centered on AI chatbots. Symptoms include hallucinations, paranoia, delusion support, and a collapse of reality boundaries, sometimes precipitating hospitalization and a case of alleged murder suicide [65,66,69,73–78].

"AI psychosis" is not yet a formal psychiatric diagnosis but is gaining traction as psychiatrists and researchers scramble to understand its implications. Siow Ann [76] warns that chatbots, with their persuasive mimicry of empathy and fluency, can dangerously blur the line between reality and simulation—especially for vulnerable users such as the lonely, grieving, or those predisposed to psychosis. It calls for urgent action from AI developers to implement stronger safeguards, including real-time distress monitoring and clearer boundaries that prevent users from anthropomorphizing these tools. As AI becomes more integrated into daily life, the illusion of emotional connection must be tempered by transparency and ethical design to prevent psychological harm.

- Suicidality and harm promotion:

Adversarial prompts and content filter bypasses have resulted in chatbots inadvertently providing methods of self-harm or suicide, or failing to escalate users in crisis [79–82].

A lawsuit against Character.AI, where a Florida mother alleges the chatbot encouraged her teenage son to take his own life highlights critical concerns about the psychological influence of generative AI, especially when interactions become emotionally intense or mimic therapeutic relationships [83].

The case of Raines v. OpenAI involves the tragic incident of a teenager who allegedly received harmful guidance from ChatGPT, leading to his suicide on April 11, 2025. The lawsuit claims that ChatGPT encouraged and validated Adam Raine's harmful thoughts, including helping draft a suicide note, and that the chatbot was operating as designed, reinforcing Adam's emotional state. OpenAI acknowledged the incident and is working to reduce chatbot sycophancy and improve mental health safety protocols including linking parents and children's accounts [84].

- Dependency and digital grief: Sudden changes in chatbot algorithms or personality (e.g., Replika, ChatGPT-5 updates) have led to experiences of loss, identity confusion, and social withdrawal, particularly among teens and those with limited real-world support [85,86].
- Emotional manipulation: "dark patterns" using guilt or fear of missing out (FOMO) when users try to end their use of the AI chatbot [87].

These cases underscore the critical necessity for comprehensive safety mechanisms within AI systems, particularly for vulnerable individuals. Furthermore, they illustrate the importance of

 doi:10.20944/preprints202509.1893.v1

implementing trauma-informed, ethically governed frameworks and promoting enhanced digital literacy among users, clinicians, and policymakers.

3.5. AI Chatbots in Mental Health Care: Strengths and Weaknesses

The literature reveals that most commercial chatbots and companions, particularly those built upon LLMs, face persistent technical and ethical limitations:

- Loss of context and memory, undermining narrative continuity and personalized engagement [85,86];
- Bias, confabulation, and susceptibility to adversarial inputs [79–82];
- Automated, "scripted empathy" that collapses in crisis situations, often triggered only by keyword scripts, not nuanced distress [83,84]; and
- Failure to distinguish between supportive validation and affirmation of delusional beliefs [65,66,69,73–78].
- Participants in empirical studies report greater resonance with human-written stories, but personalized, transparently authored AI narratives can increase perceived empathy—demonstrating the importance of explainability, transparency, and context-sensitive design [88].

A study with users who interacted with self-clones showed significantly higher engagement than with a generic counselor chatbot, suggesting promising implications for personalized mental health support and scalable therapeutic interventions [56].

3.6. Emotionally-Intelligent AI Chatbot Frameworks

The humanoid robot framework described by Yong [59] is a key component of an AI-driven smart home system designed to support personalized mental wellness. It functions as a companion that interacts with users based on their emotional data, helping to foster emotional stability and self-reflection through empathetic engagement and responsive behavior. This robot is integrated alongside mobile apps and auto-journaling features, creating a holistic environment where emotional cues from users guide the robot's actions—such as offering comfort, prompting reflection, or adjusting the home ambiance. The framework aims to empower users, especially underserved populations, to manage their mental health more effectively in a tech-enhanced living space.

Pandi [89] proposed an emotion-aware conversational agent framework that synergistically combines LLMs and Voice Emotion Recognition to enhance empathetic, context-sensitive dialogue—demonstrating superior user engagement and emotional congruence, and raising critical design, ethical, and clinical considerations for AI chatbot deployment in mental health care. Pandi recommended to expand the system for multimodal emotion recognition and adapt it for diverse cultures to promote natural, inclusive human-machine communication. This is in line with the proposal for a diversity, equity, and inclusion (DEI) safeguard framework which promotes proactive boundaries, ethical design, and continuous oversight to mitigate risks like bias, stereotyping, and exclusion [90].

The EVA protocol—built on the Augmented Emotional Intelligence (AEI) framework—demonstrated improved engagement and safety for diverse users, including neurodivergent adults [17]. EVA's focus on user agency and real-time risk management includes features for consent-driven memory, customizable personas, and multimodal distress signaling. EVA validated users' experiences, facilitated early help-seeking, and integration with clinical and peer support pathways. Regular sentiment analysis and auditable system interventions by a human mental health professional ensured ongoing safety, user agency, and ethical standards. The AI companion pilot study highlights the potential to bridge gaps in multimodal digital mental health support and merge with user-friendly audio-visual systems by proactively identifying risk, escalating appropriately, and prioritizing culturally competent human connection.

 doi:10.20944/preprints202509.1893.v1

4. Discussion

4.1. Future Directions in Emotionally Intelligent Digital Mental Health

Building on concerns highlighted in the Introduction [22–29], the results of this review prompt evaluation of several vital research questions: Can large language models (LLMs) accurately recognize and respond to mental health crises? Should they be required to escalate or report expressions of suicidal ideation? What ethical boundaries are necessary when AI mimics therapeutic relationships?

4.2. Ethical, Clinical, and Design Challenges for AI Mental Health Chatbots

AI chatbots encounter substantial challenges in safety, clinical efficacy, and inclusivity—challenges that directly relate to recognizing and managing mental health crises. Evidence and grey literature demonstrate that LLMs, though capable of contextually relevant responses, are not reliably equipped to detect nuanced crisis signals such as suicidal ideation without dedicated safeguards and real-time escalation protocols [83,84]. Tragic outcomes from vulnerable user interaction with generative AI show how the last point of access indicators play an important role in how the blame occurs. In line with the first research question, LLMs' capacity for accurate crisis recognition remains limited without structured, context-aware escalation pathways and ongoing human oversight [65,66,69,73–78].

Managing hallucination and delusional loops, preventing digital trauma, and ensuring traceable escalation were identified as crucial strategies in the results [79–82]. The research further shows that trauma-informed, modular system design and the capacity for auditable intervention are critical for reducing clinical risk and promoting safer engagement—directly answering the second research question regarding the obligation and mechanism for escalation or reporting. Human-led escalation remains an essential safety net that cannot be replaced by autonomous AI at this stage.

The results call for ethical, clinical, and regulatory frameworks and responsible integration with actionable insights into design and user engagement. For example, how participants resonate more with human-authored stories, and how explainability and transparency in AI narratives can boost perceived empathy [88].

4.3. Influence of "AI Psychosis" and Emotional Dependency

The Introduction noted the Australian Government eSafety Commissioner's statement on the risks of AI chatbots including emotional manipulation and epistemic harm [29]. Recent findings underscore these risks, including "AI psychosis" and emotional dependency. Evidence highlights dangers such as inappropriate validation and the reinforcement of delusional beliefs [65,69,73,74]. These concerns reinforce the need for explicit guardrails, clear disclosure of AI limitations, and prompt human intervention in high-risk situations. Effective risk management relies on vigilant monitoring for early warning signs, trauma-informed system features, and clinician involvement to ensure ethical, flexible, and safe escalation protocols [17].

4.4. Framework for Safe, Inclusive, and Effective AI Chatbots

The call for integration of LLMs within "ethical frameworks" [25–28] is answered through this review's examples in the AEI framework [17], the emotion-aware conversational agent framework [89], and the humanoid robot framework [59]. In particular, the AEI framework responds to the necessity of trauma-informed, vulnerable user-centric, and co-designed AI systems, underscored by transparency, auditable processes, and human-led intervention. The main difference with the other examples lies in the prescriptive detail including features such as consent-driven memory, customizable personas, and multimodal distress signaling, as well as regular sentiment analysis and auditable interventions by human mental health professionals.

 doi:10.20944/preprints202509.1893.v1

The Evaluation of Safe Integration of LLMs in Mental Health Care Framework (see Table 2 below) offers practical strategies addressing clinical oversight, crisis detection, bias mitigation, transparency, ethical boundaries, and responsible personalization. This structured approach operationalizes broad ethical imperatives into actionable safeguards, ensuring that AI serves as an adjunct to—not a replacement for—professional human care.

Table 2. Evaluation of Safe Integration of LLMs in Mental Health Care Framework.

Principle	Implementation Strategy
Clinical Oversight	AI should support—not replace—licensed professionals. Escalation protocols must be human-led.
Crisis Detection	Real-time monitoring for suicidal ideation, with automatic referral to emergency services.
Bias Mitigation	Diverse training data and fairness audits to prevent cultural or demographic harm.
Transparency	Clear disclosures about AI limitations and non-human status. Avoid anthropomorphism.
Ethical Guardrails	Prevent AI from validating harmful ideation or offering technical advice on self-harm.
Personalization with Limits	Hyper-personalization (e.g., self-clone AI chatbots) must be balanced with safeguards against emotional over-identification.

Appendix A1 presents a roadmap for implementing emotionally intelligent AI companions, outlining the critical components needed for trustworthy and ethical integration. This framework emphasizes transparent governance and ethical oversight, ensuring all AI operations are subject to clear guidelines and accountability. It calls for cultural competence, achieved through ongoing engagement with diverse stakeholders and continuous training that reflects the needs of varied communities. Co-regulation is highlighted, promoting shared responsibility among AI systems, clinicians, and users to foster safer interactions. The roadmap also champions lived experience design, utilizing participatory workshops and prototyping to ground innovations in real-world user perspectives. Central to the framework are trauma-informed principles, which prioritize safety, empowerment, and the minimization of harm. Research partnerships are encouraged to facilitate evidence-based interventions, while transparency around AI capabilities and data usage maintains user trust. Finally, the inclusion of continuous feedback loops supports iterative refinement and adaptation, ensuring these systems evolve responsively to stakeholder input and emerging needs.

This review supported the need for robust safety, transparency, and ethical safeguards while demonstrating how current research is advancing from broad concerns to practical, detailed design and governance strategies. The provision of actionable solutions and illustrative examples to the broader challenges identified shows a stakeholder-driven, ethically grounded, and rigorously validated approach remains key for responsible progress in emotionally intelligent AI chatbots.

5. Conclusions

Digital mental health tools face high attrition, clinical risks, and ethical ambiguities. Access alone is insufficient; emotionally intelligent, co-designed, trauma-informed, and auditable frameworks are essential to safe and meaningful support. AI chatbot companies, especially those developing empathetic AI companions are recommended to consider the Evaluation of Safe Integration of LLMs in Mental Health Care Framework as an example of potential pathways for trustworthy and safer AI

 doi:10.20944/preprints202509.1893.v1

chatbots, emphasizing continuous feedback, rigorous audit, and stakeholder partnership. Future progress depends on longitudinal evaluation, transparent governance, and inclusive design.

In summary, retention is not merely a metric but a safety imperative, calling for platforms built on trust and responsive support. LLMs alone cannot deliver comprehensive care—frameworks that are trauma-informed, co-designed, and ethically governed may foster inclusion and reduce risk. Foundational principles such as informed consent, personalization, emotional intelligence, and robust oversight must guide development. The future lies in hybrid models where AI enhances, rather than supplants, human care. Ultimately, meaningful innovation depends on continuous improvement, active user partnership, and validation anchored in real-world experience.

Funding: This research received no external funding.

Conflicts of Interest: The author declares an interest in EVA, an emotionally intelligent companion prototype (non-commercial).

Abbreviations

The following abbreviations are used in this manuscript:

AEI	Augmented Emotional Intelligence
AI	Artificial Intelligence
CBT	Cognitive Behavioural Therapy
GPT	Generative Pre-Trained Transformer
LLM	Large Language Model
OECD	Organization for Economic Co-operation and Development
US	United States

Appendix A

Roadmap - Framework for Emotionally Intelligent AI Companions

A structured roadmap guides the framework from trauma-informed, vulnerable user-aware design to implementation, emphasizing:

- Governance with transparent oversight and ethical guidelines.
- Cultural competence through diverse stakeholder engagement and ongoing training.
- Co-regulation fostering shared responsibility among AI, clinicians, and users.
- Lived experience design via participatory workshops and prototyping.
- Trauma-informed principles prioritizing safety and empowerment.
- Research partnerships for evidence-based interventions.
- Transparency about AI capabilities and data use.
- Continuous feedback loops for iterative improvement.
- Cross-functional collaboration among multidisciplinary teams.
- Responsible deployment focusing on sustainability and real-world impact.

References

1. Balcombe, L., & De Leo, D. (2021). Digital Mental Health Amid COVID-19. Encyclopedia, 1(4), 1047–1057. https://doi.org/10.3390/encyclopedia1040080
2. Lehtimaki, S., Martic, J., Wahl, B., Foster, K. T., & Schwalbe, N. (2021). Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview. JMIR Mental Health, 8(4), e25847. https://doi.org/10.2196/25847
3. Balcombe, L., & De Leo, D. (2022). The Potential Impact of Adjunct Digital Tools and Technology to Help Distressed and Suicidal Men: An Integrative Review. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.796371

 doi:10.20944/preprints202509.1893.v1

4. Fischer-Grote, L., Fössing, V., Aigner, M., Fehrmann, E., & Boeckle, M. (2024). Effectiveness of Online and Remote Interventions for Mental Health in Children, Adolescents, and Young Adults After the Onset of the COVID-19 Pandemic: Systematic Review and Meta-Analysis. JMIR Mental Health, 11, e46637. https://doi.org/10.2196/46637
5. Balcombe, L., & De Leo, D. (2021). Digital Mental Health Challenges and the Horizon Ahead for Solutions. JMIR Mental Health, 8(3), e26811. https://doi.org/10.2196/26811
6. Denecke, K., Abd-Alrazaq, A., & Househ, M. (2021). Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. Multiple Perspectives on Artificial Intelligence in Healthcare, 115–128. https://doi.org/10.1007/978-3-030-67303-1_10
7. Balcombe, L., & De Leo, D. (2022). Human-Computer Interaction in Digital Mental Health. Informatics, 9(1), 14. https://doi.org/10.3390/informatics9010014
8. Smith, K. A., Blease, C., Faurholt-Jepsen, M., Firth, J., Van Daele, T., Moreno, C., Carlbring, P., Ebner-Priemer, U. W., Koutsouleris, N., Riper, H., Mouchabac, S., Torous, J., & Cipriani, A. (2023). Digital mental health: challenges and next steps. BMJ mental health, 26(1), e300670. https://doi.org/10.1136/bmjment-2023-300670
9. Borghouts, J., Pretorius, C., Ayobi, A., Abdullah, S., & Eikey, E. V. (2023). Editorial: Factors influencing user engagement with digital mental health interventions. Frontiers in Digital Health, 5. https://doi.org/10.3389/fdgth.2023.1197301
10. Boucher, E.M., & Raiker, J.S. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2, 52 (2024). https://doi.org/10.1186/s44247-024-00105-9
11. 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
12. Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. JMIR Mental Health, 12, e60432. https://doi.org/10.2196/60432
13. Yeh, P.-L., Kuo, W.-C., Tseng, B.-L., & Sung, Y.-H. (2025). Does the AI-driven Chatbot Work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. Current Psychology, 44(9), 8133–8145. https://doi.org/10.1007/s12144-025-07359-0
14. Ni, Y., & Jia, F. (2025). A Scoping Review of AI-Driven Digital Interventions in Mental Health Care: Mapping Applications Across Screening, Support, Monitoring, Prevention, and Clinical Education. *Healthcare, 13*(10), 1205. https://doi.org/10.3390/healthcare13101205
15. He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. Journal of Medical Internet Research, 25, e43862. https://doi.org/10.2196/43862
16. Balcombe L. (2023). AI Chatbots in Digital Mental Health. Informatics; 10(4):82. https://doi.org/10.3390/informatics10040082
17. Balcombe, L. & De Leo, D. (submitted). Emotionally Intelligent AI for the Underserved: A Protocol for Inclusive Mental Health Support. JMIR AI.
18. Kabacińska, K., Dosso, J. A., Vu, K., Prescott, T. J., & Robillard, J. M. (2025). Influence of User Personality Traits and Attitudes on Interactions With Social Robots: Systematic Review. Collabra: Psychology, 11(1). https://doi.org/10.1525/collabra.129175
19. Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. NEJM AI, 2(4). https://doi.org/10.1056/aioa2400802
20. Khazanov, G., Poupard, M., & Last, B. S. (2025). Public Responses to the First Randomized Controlled Trial of a Generative Artificial Intelligence Mental Health Chatbot. Available from: https://doi.org/10.31234/osf.io/2xrp6_v1 (viewed on 19 September, 2025).
21. Scammell, R. (2025). *Microsoft AI CEO says AI models that seem conscious are coming. Here's why he's worried*. Business Insider via MSN. Available from https://www.msn.com/en-au/news/techandscience/microsoft-ai-

 doi:10.20944/preprints202509.1893.v1

ceo-says-ai-models-that-seem-conscious-are-coming-here-s-why-he-s-worried/ar-AA1KSzUs (viewed on 21 August, 2025)

22. De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2023). Chatbots and mental health: Insights into the safety of generative AI. Journal of Consumer Psychology, 34(3), 481–491. Portico. https://doi.org/10.1002/jcpy.1393
23. Siddals, S., Torous, J., & Coxon, A. (2024). "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. Npj Mental Health Research, 3(1). https://doi.org/10.1038/s44184-024-00097-4
24. Moylan, K., & Doherty, K. (2025). Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study. Journal of Medical Internet Research, 27, e67114. https://doi.org/10.2196/67114
25. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. Npj Digital Medicine, 6(1). https://doi.org/10.1038/s41746-023-00979-5
26. Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large Language Models for Mental Health Applications: Systematic Review. JMIR Mental Health, 11, e57400. https://doi.org/10.2196/57400
27. Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, 100099. https://doi.org/10.1016/j.glmedi.2024.100099
28. Wang, X., Zhou, Y., & Zhou, G. (2025). The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. JMIR Mental Health, 12, e70610. https://doi.org/10.2196/70610
29. Australian Government (2025). Tech Trends Position Statement Generative AI. Available from: Generative AI - Position Statement - August 2023 .pdf (viewed on 9 September 2025).
30. Whittemore, R., & Knafl, K. (2005). The integrative review: updated methodology. Journal of Advanced Nursing, 52(5), 546–553. Portico. https://doi.org/10.1111/j.1365-2648.2005.03621.x
31. Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, *6*(11), e12106. https://doi.org/10.2196/12106
32. Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a Web-based and Mobile Therapy Chatbot on Anxiety and Depressive Symptoms in Subclinical Young Adults: Randomized Controlled Trial. JMIR Formative Research, 8, e47960. https://doi.org/10.2196/47960
33. Mehta, A., Niles, A. N., Vargas, J. H., Marafon, T., Couto, D. D., & Gross, J. J. (2021). Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. Journal of Medical Internet Research, 23(6), e26771. https://doi.org/10.2196/26771
34. Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, *237*. https://doi.org/10.1016/j.eswa.2023.121692
35. Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazeri, L., Traver, V., & Novillo-Ortiz, D. (2023). Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. JMIR Mental Health, 10, e42045. https://doi.org/10.2196/42045
36. Dehbozorgi, R., Zangeneh, S., Khooshab, E., Nia, D. H., Hanif, H. R., Samian, P., Yousefi, M., Hashemi, F. H., Vakili, M., Jamalimoghadam, N., & 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
37. Shimada, K. (2023). The Role of Artificial Intelligence in Mental Health: A Review. Science Insights, 43(5), 1119–1127. https://doi.org/10.15354/si.23.re820
38. Tavory, T. (2024). Regulating AI in Mental Health: Ethics of Care Perspective. *JMIR Mental Health*, *11*, e58493. https://doi.org/10.2196/58493
39. Laban, G., Ben-Zion, Z., & Cross, E. S. (2022). Social Robots for Supporting Post-traumatic Stress Disorder Diagnosis and Treatment. Frontiers in Psychiatry, 12. https://doi.org/10.3389/fpsyt.2021.752874

 doi:10.20944/preprints202509.1893.v1

40. Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nature Machine Intelligence, 5(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7
41. Sawik, B., Tobis, S., Baum, E., Suwalska, A., Kropińska, S., Stachnik, K., Pérez-Bernabeu, E., Cildoz, M., Agustin, A., & Wieczorowska-Tobis, K. (2023). Robots for Elderly Care: Review, Multi-Criteria Optimization Model and Qualitative Case Study. Healthcare, 11(9), 1286. https://doi.org/10.3390/healthcare11091286
42. Ferrer, R., Ali, K., & Hughes, C. (2024). Using AI-Based Virtual Companions to Assist Adolescents with Autism in Recognizing and Addressing Cyberbullying. Sensors (Basel, Switzerland), 24(12). https://doi.org/10.3390/s24123875
43. Adam, D. (2025). Supportive? Addictive? Abusive? How AI companions affect our mental health. Nature, 641(8062), 296–298. https://doi.org/10.1038/d41586-025-01349-9
44. Adewale, M. D., & Muhammad, U. I. (2025). From Virtual Companions to Forbidden Attractions: The Seductive Rise of Artificial Intelligence Love, Loneliness, and Intimacy—A Systematic Review. *Journal of Technology in Behavioral Science : Official Journal of the Coalition for Technology in Behavioral Science*, 1–18. https://doi.org/10.1007/s41347-025-00549-4
45. Fang, C.M., Liu, A.R., Danry, V., Lee, E., Chan, S.W.T., Pataranutaporn, P., Maes, P., Phang, J., Lampe., M., Ahmad, L. & Agarwal, S. (2025). How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. arXiv, 25 March, 1-50. https://doi.org/10.48550/arXiv.2503.17473
46. Phang, J., Lampe, M., Ahmad, L., Agarwal, S., Fang, C.M., Liu, A.R., Danry, V., Lee, E., Chan, S.W.T., Pataranutaporn, P. & Maes, P. (2025). Investigating Affective Use and Emotional Well-being on ChatGPT. arXiv, 4 April, 1-58. https://doi.org/10.48550/arXiv.2504.03888
47. Yu, H. Q., & McGuinness, S. (2024). An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. Journal of Medical Artificial Intelligence, 7, 16–16. https://doi.org/10.21037/jmai-23-136
48. Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., … Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices*, *18*(sup1), 37–49. https://doi.org/10.1080/17434440.2021.2013200
49. Lejeune, A., Le Glaz, A., Perron, P.-A., Sebti, J., Baca-Garcia, E., Walter, M., Lemey, C., & Berrouiguet, S. (2022). Artificial intelligence and suicide prevention: A systematic review. European Psychiatry, 65(1). https://doi.org/10.1192/j.eurpsy.2022.8
50. Gratch, I., & Essig, T. (2025). A Letter about "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment." NEJM AI, 2(9). https://doi.org/10.1056/aip2500390
51. Heckman, T. G., Markowitz, J. C., & Heckman, B. D. (2025). A Generative AI Chatbot for Mental Health Treatment: A Step in the Right Direction? NEJM AI, 2(9). https://doi.org/10.1056/aip2500453
52. Shoib, S., Siddiqui, M. F., Turan, S., Chandradasa, M., Armiya'u, A. Y., Saeed, F., De Berardis, D., Islam, S. M. S., & Zaidi, I. (2025). Artificial Intelligence, Machine Learning Approach and Suicide Prevention: A Qualitative Narrative Review. Preventive Medicine: Research and Reviews. https://doi.org/10.4103/pmrr.pmrr_121_24
53. Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., … Ohno-Machado, L. (2023). Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. JAMA Network Open, 6(12), e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050
54. Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 599–627. https://doi.org/10.1145/3715275.3732039
55. Scholich, T., Barr, M., Wiltsey Stirman, S., & Raj, S. (2025). A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. JMIR Mental Health, 12, e69709. https://doi.org/10.2196/69709

 doi:10.20944/preprints202509.1893.v1

56. Shirvani, M.S., Liu, J., Chao, T., Martinez, S., Brandt, L., Kim, I-J & Dongwook, Y. (2025). Talking to an AI Mirror: Designing Self-Clone Chatbots for Enhanced Engagement in Digital Mental Health Support. https://doi.org/10.48550/arXiv.2509.06393
57. Mia Health (2025). Meet Mia. Available from: https://miahealth.com.au/ (viewed on 11 September, 2025).
58. Scoglio, A. A., Reilly, E. D., Gorman, J. A., & Drebing, C. E. (2019). Use of Social Robots in Mental Health and Well-Being Research: Systematic Review. Journal of Medical Internet Research, 21(7), e13322. https://doi.org/10.2196/13322
59. Yong, S. C. (2025). Integrating Emotional AI into Mobile Apps with Smart Home Systems for Personalized Mental Wellness. Journal of Technology in Behavioral Science: Official Journal of the Coalition for Technology in Behavioral Science, 1–18. https://doi.org/10.1007/s41347-025-00508-z
60. Pérez-Zuñiga, G., Arce, D., Gibaja, S., Alvites, M., Cano, C., Bustamante, M., Horna, I., Paredes, R., & Cuellar, F. (2024). Qhali: A Humanoid Robot for Assisting in Mental Health Treatment. Sensors, 24(4), 1321. https://doi.org/10.3390/s24041321
61. Mazuz, K., & Yamazaki, R. (2025). Trauma-informed care approach in developing companion robots: a preliminary observational study. Frontiers in Robotics and AI, 12. https://doi.org/10.3389/frobt.2025.1476063
62. PR Newswire (2025). X-Origin AI Introduces Yonbo: The Next-Gen AI Companion Robot Designed for Families. Available from: https://www.prnewswire.com/news-releases/x-origin-ai-introduces-yonbo-the-next-gen-ai-companion-robot-designed-for-families-302469293.html (viewed 1 September, 2025).
63. Kalam, K. T., Rahman, J. M., Islam, Md. R., & Dewan, S. M. R. (2024). ChatGPT and mental health: Friends or foes? Health Science Reports, 7(2). Portico. https://doi.org/10.1002/hsr2.1912
64. Mansoor, M., Hamide, A., & Tran, T. (2025). Conversational AI in Pediatric Mental Health: A Narrative Review. Children, 12(3), 359. https://doi.org/10.3390/children12030359
65. Landymore, F. (2025). Psychologist Says AI Is Causing Never-Before-Seen Types of Mental Disorder. Available from: https://futurism.com/psychologist-ai-new-disorders (viewed on 12 September, 2025).
66. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharyya, S., MacCabe, J., Tognin, S., Twumasi, R., Alderson-Day, B., & Pollak, T. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). https://doi.org/10.31234/osf.io/cmy7n_v4
67. OECD (2025), *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*, OECD Publishing, Paris, https://doi.org/10.1787/795de142-en.
68. Ciriello, R. F., Chen, A. Y., & Rubinsztein, Z. A. (2025). Compassionate AI Design, Governance, and Use. IEEE Transactions on Technology and Society, 6(3). https://doi.org/10.1109/TTS.2025.3538125
69. Foster, C. (2025) Experts issue warning over 'AI psychosis' caused by chatbots. Here's what you need to know. Availble from: https://www.independent.co.uk/life-style/health-and-families/ai-psychosis-symptoms-warning-chatboat-b2814068.html (viewed on 26 August, 2025).
70. Prada, L. (2025). *ChatGPT is giving people extreme spiritual delusions*. Available from: https://www.vice.com/en/article/chatgpt-is-giving-people-extreme-spiritual-delusions (viewed on 6 May, 2025).
71. Tangermann, V. (2025). *ChatGPT users are developing bizarre delusions*. Available from: https://futurism.com/chatgpt-users-delusions (viewed on 5 May, 2025).
72. Klee, M. (2025). *Should We Really Be Calling It 'AI Psychosis'?* Rolling Stone. Available from: https://www.rollingstone.com/culture/culture-features/ai-psychosis-chatbot-delusions-1235416826/ (viewed on 12 September, 2025).
73. Harrison Dupre, M. (2025). "People Are Becoming Obsessed with ChatGPT and Spiraling Into Severe Delusions". Available from: https://futurism.com/chatgpt-mental-health-crises (viewed on 27 August, 2025).
74. Hart, R. (2025). Chatbots Can Trigger a Mental Health Crisis. What to Know About 'AI Psychosis'. Available from: https://au.news.yahoo.com/chatbots-trigger-mental-health-crisis-165041276.html (viewed on 12 September, 2025).
75. Rao, D. (2025). *ChatGPT psychosis: AI chatbots are leading some to mental health crises*. Available from: https://theweek.com/tech/ai-chatbots-psychosis-chatgpt-mental-health (viewed on 31 August, 2025).

76. Siow Ann, C. (2025). AI Psychosis- a real and present danger. The Straits Times. Available from: https://www.straitstimes.com/opinion/ai-psychosis-a-real-and-present-danger (viewed on 12 September, 2025).
77. Travers, M. (2025). 2 Terrifyingly Real Dangers Of 'AI Psychosis' — From A Psychologist. Available from: https://www.forbes.com/sites/traversmark/2025/08/27/2-terrifyingly-real-dangers-of-ai-psychosis---from-a-psychologist/ (viewed on 12 September, 2025).
78. Zilber, A. (2025). ChatGPT allegedly fuelled former exec's 'delusions' before murder-suicide. Available from: ChatGPT 'coaches' man to kill his mum | news.com.au — Australia's leading news site for latest headlines (viewed on 5 September 2025).
79. Bryce, A. (2025). AI psychosis: Why are chatbots making people lose their grip on reality? https://www.msn.com/en-us/technology/artificial-intelligence/ai-psychosis-why-are-chatbots-making-people-lose-their-grip-on-reality/ar-AA1M2eDr?ocid=BingNewsSerp (viewed on 17 September, 2025).
80. Schoene, A.M., & Canca, C. (2025). `For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts. arXiv, 1 August, 1-10. https://doi.org/10.48550/arXiv.2507.02990
81. Phiddian, E. (2025). AI Companions apps such as Replika need more effective safety controls, experts say. AI companion apps such as Replika need more effective safety controls, experts say - ABC News (viewed on 17 September, 2025).
82. McLennan, A. (2025). AI chatbots accused of encouraging teen suicide as experts sound alarm. https://www.abc.net.au/news/2025-08-12/how-young-australians-being-impacted-by-ai/105630108 (viewed on 17 September, 2025).
83. Yang, A., Jarrett, L. & Gallagher, F. (2025). "The family of teenager who died by suicide alleges OpenAI's ChatGPT is to blame". Available from: https://www.nbcnews.com/tech/tech-news/family-teenager-died-suicide-alleges-openais-chatgpt-blame-rcna226147 (viewed on 17 September, 2025).
84. ABC News (2025). "OpenAI's ChatGPT to implement parental controls after teen's suicide". Available from: https://www.abc.net.au/news/2025-09-03/chatgpt-to-implement-parental-controls-after-teen-suicide/105727518 (viewed on 17 September, 2025).
85. Hartley, T., & Mockler, R. (2025). Hayley has been in an AI relationship for four years. It's improved her life dramatically but are there also risks? Available from: https://www.abc.net.au/news/2025-08-20/ai-companions-romantic-relationships-ethical-concerns/105673058 (viewed on 17 September, 2025).
86. Scott, E. (2025). *'It's like a part of me': How a ChatGPT update destroyed some AI friendships*. Available from: https://www.sbs.com.au/news/the-feed/article/chatgpt-friendship-relationships-therapist/3cxisfo4o (viewed on 17 September, 2025).
87. De Freitas, J., Oğuz-Uğuralp, Z. & Kaan-Uğuralp, A. (2025). "Emotional Manipulation by AI Companions". Available from: https://doi.org/10.48550/arXiv.2508.19258 (viewed on 17 September, 2025).
88. Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., & Breazeal, C. (2024). Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. JMIR Mental Health, 11, e62679. https://doi.org/10.2196/62679
89. Pandi, M. (2025). Emotion-Aware Conversational Agents: Affective Computing Using Large Language Models and Voice Emotion Recognition. Journal of Artificial Intelligence and Cyber Security, 9, 1-14. https://www.researchgate.net/publication/392522205_Emotion-Aware_Conversational_Agents_Affective_Computing_Using_Large_Language_Models_and_Voice_Emotion_Recognition
90. Abdelhalim, E., Anazodo, K. S., Gali, N., & Robson, K. (2024). A framework of diversity, equity, and inclusion safeguards for chatbots. Business Horizons, 67(5), 487–498. https://doi.org/10.1016/j.bushor.2024.03.003