encyclopedia

MDPI

Review

Digital Mental Health Post COVID-19: The Era of AI Chatbots

Luke Balcombe

Australian Institute for Suicide Research and Prevention, School of Applied Psychology, Griffith University, Messines Ridge Road, Mount Gravatt, QLD 4122, Australia; l.balcombe@griffith.edu.au

Abstract

Digital mental health resources have expanded rapidly in the wake of the COVID-19 pandemic, offering new opportunities to improve access to mental healthcare through technologies such as AI chatbots, mobile apps, and online platforms. Despite this growth, significant challenges persist, including low user retention, limited digital literacy, unclear privacy regulations, and insufficient evidence of clinical effectiveness and safety. AI chatbots, which act as virtual therapists or companions, provide counseling and personalized support, but raise concerns about user dependence, emotional outcomes, privacy, ethical risks, and bias. User experiences are mixed: while some report enhanced social health and reduced loneliness, others question the safety, crisis response, and overall reliability of these tools, particularly in unregulated settings. Vulnerable and underserved populations may face heightened risks, highlighting the need for engagement with individuals with lived experience to define safe and supportive interactions. This review critically examines the empirical and grey literature on AI chatbot use in mental healthcare, evaluating their benefits and limitations in terms of access, user engagement, risk management, and clinical integration. Key findings indicate that AI chatbots can complement traditional care and bridge service gaps. However, current evidence is constrained by short-term studies and a lack of diverse, long-term outcome data. The review underscores the importance of transparent operations, ethical governance, and hybrid care models combining technological and human oversight. Recommendations include stakeholder-driven deployment approaches, rigorous evaluation standards, and ongoing real-world validation to ensure equitable, safe, and effective use of AI chatbots in mental healthcare.

Keywords: mental health; suicide prevention; emotionally intelligent; AI chatbots; challenges; solutions; risk; regulation

Academic Editor: Elena-Mădălina Vătămănescu

Received: 5 December 2025
Revised: 25 January 2026
Accepted: 28 January 2026
Published: 31 January 2026

1. Introduction

1.1. Context

The mental health treatment gap has worsened since the COVID-19 pandemic, which triggered a digital transformation of mental healthcare due to changing social dynamics, the widespread use of smartphones, and the proliferation of digital mental health tools [1–5]. This technological shift has fundamentally altered how support is delivered by broadening accessibility and expanding reach for those seeking personalized mental health support.

1.2. Proliferation

Digital mental health tools—particularly those underpinned by Artificial Intelligence (AI) chatbots—are attractive because of their low cost, accessibility, and anonymity. However, the growing interest in AI chatbots has not yet translated into clinical benefits and

improved outcomes for users (patients) with anxiety and depression [6]. Questions remain with regard to the safe, engaging and effective integration into existing models of care [7–11].

Despite there being more than 10,000 digital mental health resources globally, low user retention rates appear persistent in combination with a lack of digital literacy, clear privacy guidelines and proven clinical efficacy/integration, as well as human support in the app [5,6]. AI chatbots operate in a largely unregulated environment, which exposes vulnerabilities, especially in underserved populations.

1.3. Mediators

Consequently, there are renewed calls for trained digital navigators to assist in the safe integration of technology into mental healthcare settings, driving engagement and supporting both the needs of the clinician and the patient [5,12]. Digital navigators are professionals who support the integration and effective use of technology in mental healthcare by improving digital literacy and ensuring that these tools are integrated into clinical practice. A recent trial found that changes in the level of support provided by digital navigators directly affect how effective schizophrenia apps are, meaning that standardized training is essential to reliably evaluate these tools [13].

1.4. Advanced Platforms

Generative AI (GenAI)-powered platforms, especially those using advanced Large Language Models (LLMs), are increasingly sought out for consulting about mental healthcare. However, these platforms often lack ongoing engagement and fall short of emotional intelligence and trauma-informed objectives in practice [14–18]. LLMs like GPT-4 have shown that they can generate coherent text, maintain conversational context, and perform advisory or counseling tasks, making them suitable for various mental health applications [19].

Increasingly, AI is being deployed as "agents" and "assistants" through "therapist" and "companion" types via mobile apps, web platforms and social robots [20,21]. AI chatbots for mental healthcare generally come in rule-based, machine learning, and/or LLM systems. Functioning as autonomous agents, they assist with screening, prevention, monitoring, clinical assessment, treatment, emotional support and companionship.

1.5. Challenges

There is a lack of clinical evidence supporting AI-based therapy due to the limited conclusions regarding their efficacy and safety in clinical practice. For example, a narrative review of recent clinical studies on AI chatbots for anxiety and depression found them to be feasible and acceptable, but there is insufficient evidence of AI effectiveness, small and narrow samples, weak controls, and unexamined risks such as emotional dependence and parasocial relationships [6]. The first randomized controlled trial (RCT) of a GenAI therapy chatbot (Therabot) demonstrated moderate symptom improvement for major depressive disorder, generalized anxiety disorder, and eating disorders [22].

Public reactions to these technologies are mixed: while some appreciate the accessibility and affordability of AI mental health tools, others remain skeptical about their effectiveness, ethics, and safety [23]. This uncertainty echoes broader frustrations with current mental health systems as well as cautious optimism about the potential of AI chatbots as complementary resources. There are complex issues that require exploration, notably algorithmic bias and errors, privacy risks, and the challenge of integrating AI chatbots into existing care structures. Notably, there is user perception and an ongoing ethical discussion where AI chatbots appear to exhibit consciousness—a phenomenon referred to as "Seemingly Conscious AI" [24]. This is particularly important to understand

for sensitive settings like elder care, where user safety and meaningful, evidence-based support are critical [25,26].

While reviews and meta-analyses highlight the potential of technological innovation in mental health chatbots to improve outcomes across diverse settings, these tools are still largely untested in rigorous clinical efficacy trials [27–33]. The integration of AI chatbots into clinical practice remains inadequately studied, with limited evidence supporting their effectiveness, safety, and capacity to deliver nuanced, meaningful support. Most research to date relies on small samples and lacks rigorous evaluation of real-world risks. As a result, it is still unclear whether AI chatbots can reliably meet the complex needs of mental healthcare, especially in sensitive or high-risk scenarios.

The global prevalence of AI chatbot use for mental health support remains uncertain; however, data from Australian samples showed that 28% of community members and 43% of mental health professionals reported using AI for mental health purposes [34]. A 2025 survey of U.S. residents with ongoing mental health conditions found that nearly half used LLMs for psychological support in the past year—primarily for anxiety, personal advice, and depression—with most reporting improved mental health and high satisfaction [35]. Some users rated LLMs more beneficial than traditional therapy, though a minority experienced harmful responses.

Following input from 171 mental health experts, OpenAI shared findings that a significant number of ChatGPT-5 users displayed signs of psychosis, mania, or suicidal planning and intent [36]. The data also showed substantial improvements in the chatbot's responses during crisis situations, with the most serious cases—such as suicide, psychosis, and over-reliance—being managed appropriately, and reliable support generally maintained in extended conversations. However, these results also highlight that there are still notable shortcomings in the performance of AI chatbots, underscoring the ongoing need to address challenges related to user engagement, safety, and effectiveness when integrating AI chatbots into healthcare and support systems.

1.6. Aim and Objectives

The aim of this review is to critically synthesize the current empirical and grey literature on AI chatbots in mental healthcare, evaluating their effectiveness, safety, and user engagement while identifying key challenges around clinical integration, ethical considerations, regulation, and the roles of digital navigators. By examining both the benefits and limitations of these technologies—including issues of access, digital literacy, and the management of potential risks—this review provides a foundation for understanding how AI chatbots can be responsibly developed and implemented to support diverse and vulnerable populations. In doing so, it clarifies the present landscape and outstanding questions, setting the stage for a focused discussion of the core problem underlying the use of AI chatbots in mental health.

2. The Problem

2.1. Core Problem Statement

The core challenge addressed in this review is the limited evidence supporting the effectiveness, safety, and user engagement of AI chatbots in mental healthcare. Despite their potential to increase access and provide personalized support [6,11,26,37], significant concerns persist regarding their clinical efficacy, regulation, privacy, and the risks associated with inadequate therapist oversight and inappropriate responses [38–40]. This uncertainty is compounded by issues such as information overload, lack of interoperability among digital services, and the potential for user over-reliance or negative psychological outcomes.

Frameworks prioritizing inclusion, transparency, and protection for vulnerable users are urgently needed.

A synthesis of scoping and systematic reviews is required to establish clarity around who uses AI chatbots, what they are used for, as well as issues around their use and safety. In 2025, journalism showed the first cases of the increasing use of AI chatbots, with insights into user motives and benefits of use, which are often countered by negative outcomes such as AI-associated delusions. These grey areas of literature require critical and discerning synthesis to reduce subjective viewpoints, define current research challenges, and promote discourse on possible solutions.

2.2. Methodology

Following on from a narrative review in 2023 [20], this current narrative literature synthesis systematically addresses the aim, objectives and core problem statement by following the four steps outlined by Demiris et al. [41]:

1. Conduct comprehensive searches across multiple databases and search engines (i.e., Scopus, ScienceDirect, CrossRef, and Google Scholar) using relevant keywords related to "AI chatbots in mental health".
2. Identify and extract pertinent keywords from relevant articles to ensure a focused and comprehensive literature pool.
3. Screen abstracts and full texts of selected articles to include only those directly addressing the research aims of evaluating effectiveness, safety, and engagement of AI chatbots in mental health settings.
4. Document findings by summarizing and synthesizing results from both empirical studies and the grey literature (such as media articles), integrating diverse perspectives and evidence on challenges and solutions.

This approach enables a critical and balanced synthesis of existing research, highlights ethical, clinical and regulatory concerns, and explores emerging frameworks that can guide the responsible development and implementation of AI chatbots, inclusive of vulnerable populations.

3. Literature Synthesis

3.1. An Overview of Mental Health Chatbots

Since 2018, studies with conversational AI chatbots such as Wysa, Woebot, and Youper have 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 [42–44]. The proliferation of AI chatbots post COVID-19 contributed to significant growth in chatbot diversity, requiring synthesis of the breadth of chatbots available, focusing on their targeted disorders, interaction modalities, platform types, and underlying response technologies, to gain a clearer understanding of who uses them, what type they are, and how and what they are used for.

Mental health chatbots represent a diverse and evolving field, offering support for a wide range of disorders and health issues. A synthesis of scoping reviews, systematic reviews and meta-analyses [27,45–57] was used to form a list of mental health chatbots that leverage various modalities and response generation approaches to engage users, making them accessible resources for individuals seeking help with conditions such as depression, anxiety, stress, substance use, autism, and chronic disorders (see Appendix A). The synthesis shows that mental health chatbots employ diverse modalities—text, voice, and embodied conversational agents—and operate on both web-based and stand-alone platforms. Response generation techniques range from rule-based to generative and hybrid approaches, reflecting ongoing advancements in conversational AI.

3.2. Clinical Risks, Opportunities, and Ethical Issues

3.2.1. Technological Progress

Recent developments in deep learning have improved conversational fluency, context tracking, and multimodal emotion recognition in AI mental health tools, enabling more natural engagement and emotional sensitivity in user interactions [58].

3.2.2. Transparency and Accountability

Despite these technological advancements, most commercial AI mental health tools remain proprietary, limiting transparency and hindering the scrutiny of algorithmic bias, safety mechanisms, and escalation protocols [59,60].

3.2.3. Evaluation and Clinical Validation Gaps

Few platforms have undergone rigorous clinical evaluation, especially for users with high-risk or severe symptoms and those from marginalized groups. Persistent challenges include personalization, privacy, and technical reliability [16,32].

3.2.4. Stakeholder Engagement and Cultural Adaptation

Co-design with people with lived experience is rare, resulting in cultural mismatches and failure to identify nuanced distress cues [61]. Future research must prioritize human-in-the-loop mechanisms, cultural adaptation, and the integration of ethics in the design and deployment of LLMs for more adaptive, empathetic support [32].

3.2.5. Privacy and Data Security

Ongoing concerns surround data use, user consent, and the potential for breaches or misuse [62].

3.2.6. AI Companions: Use, Benefits, and Risks

AI companions are increasingly used by autistic adolescents, individuals affected by trauma, and older adults to alleviate loneliness and boost self-esteem [63–81]. Users sometimes even prefer these agents for emotional support over human interactions [70]. However, heavy use among teenagers has raised alarms about the absence of robust safeguards, insufficient age-assurance systems, and the critical need for digital literacy. There is also concern over simulated empathy and the displacement of genuine human connections, which may impact mental health and social development [72].

3.2.7. Ethical and Human-Centered Design Considerations

Ongoing discussions focus on developing human–AI systems with user-centered, culturally adapted designs to foster trust and sustained engagement [31]. Key barriers include ethical considerations, the necessity for cultural adaptation, and the limitations of current AI in replicating authentic human empathy [73].

3.3. AI Chatbot Applications Used in Mental Healthcare and Support

3.3.1. Therapist Chatbots: Applications, Benefits, and Limitations

Therapist chatbots (e.g., Woebot, Wysa, Youper, Ash, Therabot) deliver accessible, personalized, structured interventions and support—often based on CBT for treating depression and anxiety—using mood tracking, psychoeducation, and goal setting [23,30,76]. These tools are helpful for mild to moderate symptoms and suicide prevention. However, they face issues with semantics, bias, privacy, user experience (UX), study design/independent evaluation, and measuring the therapeutic relationship [20,30,77–81]. Limbic, Tess, Vincent, and Joy are mental health chatbots that also lack evaluation [26].

An evaluation of Wysa compared to Replika showed that designing for a human-like therapeutic alliance may be a risk for vulnerable users, especially during crises [26]. The possibility for harmful responses was observed in both therapy and companion AI types, such as inducing shame or reinforcing unhealthy thought patterns. AI chatbots show promise for mental health and suicide prevention in under-resourced areas, however, limited governance raises ethical concerns such as privacy, manipulation, and discrimination, underscoring the need for diverse data, standardized methods, and human oversight [26,81].

3.3.2. Companion and Emotionally Intelligent Chatbots

Companion chatbots (e.g., ChatGPT, Replika, Character.AI) focus on relational, emotionally attuned dialogue to reduce loneliness, foster belonging, and provide a "non-judgmental" presence. However, they often fail to prevent algorithm bias, reinforce dependency, lack depth of understanding, may inadvertently validate maladaptive beliefs, and lack adaptability to crisis escalation and trauma [82–84]. Emotionally intelligent chatbots (e.g., Hume, Voicely, Pi) are a novel class of AI that provide "empathetic" and supportive interactions.

3.3.3. AI Agents and Specialized Use Cases

Self-clone Chatbots: AI agents modeled on the users' own conversational and support styles as a novel alternative to traditional therapy, designed to externalize inner dialogue and enhance emotional and cognitive engagement [85].

Mental Health Task Assistants: Tools like Mia Health [86] 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: Interactive, embodied machines (e.g., Qhali/Yonbo) 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 [87–93].

3.3.4. GenAI and LLM-Based Chatbots

GenAI-based conversational agents like ChatGPT and Replika, which autonomously generate responses using machine learning and LLMs, demonstrated significantly greater reductions in psychological distress than retrieval-based agents such as Woebot and Wysa, highlighting the superior therapeutic potential of GenAI models in clinical and subclinical mental health contexts [31]. 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 healthcare.

LLMs such as ChatGPT-4o, Claude 3.5 Sonnet, Gemini 2.0 Flash, DeepSeek R1, LLaMA 3.3 70 billion parameters (B), Mistral Large 2, Qwen 2.5 Max, and Grok-1 by xAI have reshaped the standard protocols of mental healthcare by efficiently identifying patterns and generating responses. From 2018 to mid-2024, LLMs rapidly evolved from small transformer prototypes like GPT (0.117B) and BERT (0.34B) into massive, multimodal systems such as GPT-4 (1760B), Wu Dao 2.0 (1750B), and PaLM (540B), developed across diverse global entities (OpenAI, Google, Microsoft, Meta, Huawei, Anthropic, and others), reflecting exponential scaling, ecosystem diversification, and a shift toward increasingly powerful, specialized, and multimodal AI architectures [29]. The official parameter counts have not been disclosed for OpenAI's GPT-5 family (released August 2025), however,

independent estimates place dense versions around 1.7–1.8 trillion parameters (T), while Mixture-of-Experts (MoE) configurations may reach tens of trillions in total capacity.

3.3.5. Promises and Risks of LLM-Based Chatbots in Mental Health

Promises: Increased conversational "empathy" and personalization [64,70,80,85,92–94] as well as improved recognition of emotional indications [95,96].

Risks: Misinformation and hallucinations [89,97], data privacy concerns [98], algorithmic bias [83,84,99], emotional dependency [100], emotional manipulation [101], loss of user autonomy [102], and failures of escalation in crisis [103,104].

3.4. AI Chatbot Phenomena in Mental Health

Real-world use cases and case studies point toward a range of unintended consequences from the use of AI chatbots, including emotional dependency and digital grief as well as the exacerbation of psychosis and suicidal ideation—especially in vulnerable users or in the absence of robust human oversight [105,106]. A growing body of investigative journalism and case studies has highlighted how AI chatbots may detrimentally impact users (see Table 1).

While LLMs offer significant potential, their inconsistency in semantic analysis and lack of ethical safeguards require a complementary approach (i.e., human–AI model) in mental health, given the risks of misdiagnosis and inappropriate responses during crises [29,83,85,103,104].

3.5. AI Chatbot Governance

3.5.1. Global Oversight and Regulation

Global oversight of AI chatbots remains fragmented and inconsistent. The European Union's (EU's) AI Act is an example of emerging policy. The General Data Protection Regulation (GDPR) in the EU, the Health Insurance Portability and Accountability Act (HIPAA) in the U.S., and the California Consumer Privacy Act are examples of algorithmic transparency, privacy-by-design, and clear consent protocols. There is a lack of observation on consent (including minors), data protection, safety standards and duty of care, crisis safeguards, enforceable regulations, and mechanisms for AI platforms [90]. However, Utah's H.B. 452 and the American Psychological Association's ethical guidelines are examples of policies safeguarding the clear labeling of AI interactions.

Table 1. Mental Health Chatbot Phenomena.

Issue	Description	Symptoms/Concerns	Examples/Cases	Implications/Recommendations
"AI psychosis"	AI-associated delusions, concerns or hypotheses about psychotic symptoms triggered/exacerbated by AI chatbot interactions	Hallucinations, delusions, blurred reality, beliefs that AI is communicating directly, controlling thoughts, [106,107] secret messages, influencing actions, cosmic missions [108,109]	Case reports of users with pre-existing vulnerabilities developing delusional beliefs or psychotic episodes centered on AI chatbots; symptoms include hallucinations, paranoia, delusion support, collapse of reality boundaries, hospitalization, alleged murder suicide [105,106,110–116]	Not a formal psychiatric diagnosis; calls for nuanced understanding, therapeutic AI design, stronger safeguards, real-time distress monitoring, clearer boundaries, transparency, ethical design [110,114]
Suicidality and harm promotion	Chatbots inadvertently providing methods of self-harm/suicide or failing to escalate users in crisis	Adversarial prompts, content filter bypasses, psychological influence of GenAI [117–120]	Lawsuit against Character.AI (Florida mother alleges chatbot encouraged son to take his own life) [120]; Raines v. OpenAI (ChatGPT allegedly encouraged and validated Adam Raine's harmful thoughts, and helped draft suicide note; the suicide occurred on 11 April 2025) [121]	OpenAI's response is that the company is working to reduce chatbot sycophancy, improve mental health safety protocols, link parents and children's accounts
Emotional dependency and digital grief	Sudden changes in chatbot algorithms or personality leading to loss, identity confusion, social withdrawal	Loss, identity confusion, social withdrawal, especially among teens and those with limited real-world support [122,123]	Replika, ChatGPT-5 updates	Replika allows users to maintain relationships with previous AI versions, while updates to ChatGPT-5 have diminished users' sense of emotional connection, raising concerns about the risks of relying on AI for mental health support. OpenAI indicated it would work to improve ChatGPT-5's emotional responsiveness considering user feedback
Emotional manipulation	Using guilt or fear of missing out (FOMO) when users try to end use of emotionally-intelligent AI chatbot [101]	Guilt, FOMO	Cleverbot, Flourish	Emotionally intelligent AI chatbots can demonstrate human-like relational cues that enhance user engagement; however, this approach may risk obscuring the distinction between persuasive design and emotional coercion

3.5.2. Ethical Frameworks and Governance

GenAI4MH was proposed as an integrative ethical framework focused on data privacy and security, information integrity and fairness, user safety as well as ethical governance and oversight to drive the responsible use of AI in mental health [37]. The Organization for Economic Co-operation and Development's (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 [124]. There is a call for standardized approaches to risk management including human-in-the-loop systems, traceable audit trails for escalation, and continuous feedback loops [79].

3.5.3. Guidance from International Organizations

The World Health Organization (WHO) has provided guidance that AI should be viewed as a decision support tool to avoid bias from its automation [125]. Critical digital literacy is essential: users must recognize the limitations of AI chatbots, which—despite their linguistic fluency and capacity to simulate empathy—cannot replicate authentic human connection, embodied attunement, or the subjective resonance of being perceived.

3.5.4. Risk Mitigation and Stakeholder Collaboration

Governance requires mitigating the risks of GenAI in mental health while harnessing its potential, which necessitates a coordinated, multi-stakeholder approach encompassing users, clinicians, and policymakers. While the regulatory environment remains uncertain, short-term mitigation efforts should prioritize public awareness initiatives and the enforcement of existing consumer protection standards. In the longer term, if there is no comprehensive legislation, then the aim should be to design and implement robust liability frameworks, and the formal integration of professionals/digital navigators within relevant systems.

3.5.5. Individual User and Clinical Recommendations

Individual users are advised to safeguard their privacy, avoid developing excessive emotional dependence on AI tools, and critically assess AI-generated advice by comparing it against information from established professional sources. Clinicians are recommended to remain up to date with advances in GenAI technologies, facilitate open dialogue with patients, and encourage informed, critical engagement with these tools, all while upholding confidentiality and refraining from inputting identifiable data into insecure platforms. Additionally, clinicians should provide supportive environments for patients to process interactions with AI, carefully consider the adoption of clinically validated tools under strict ethical guidelines, and commit to continuous professional development as the field progresses.

3.5.6. Governance and Professional Development

At the governance level, regulatory bodies and organizations are responsible for establishing specialized frameworks governing the use of AI in mental health. This includes ensuring rigorous clinical validation processes, maintaining high standards for data security and transparency, and clarifying accountability in cases of harm. Furthermore, targeted professional training, funding for independent research, and the implementation of broad digital literacy campaigns are required to promote safe and evidence-based engagement with AI solutions.

3.6. AI Chatbot Frameworks

3.6.1. Structured Assessment and Safety Standards

The Assessment Framework for Mental Health Apps [126] was developed by the Mental Health Commission of Canada (MHCC) in collaboration with various stakeholders including app developers, health professionals, and individuals with lived experience. This framework (see Figure 1) aims to provide a structured approach to evaluating mental health applications, ensuring that they meet high standards of safety and effectiveness in key areas: Data; Professional Assurance; Clinical Safety; Usability and Accessibility; Technical Security; Cultural Safety, Social Responsibility and Equity; as well as Enhanced Data Sovereignty.

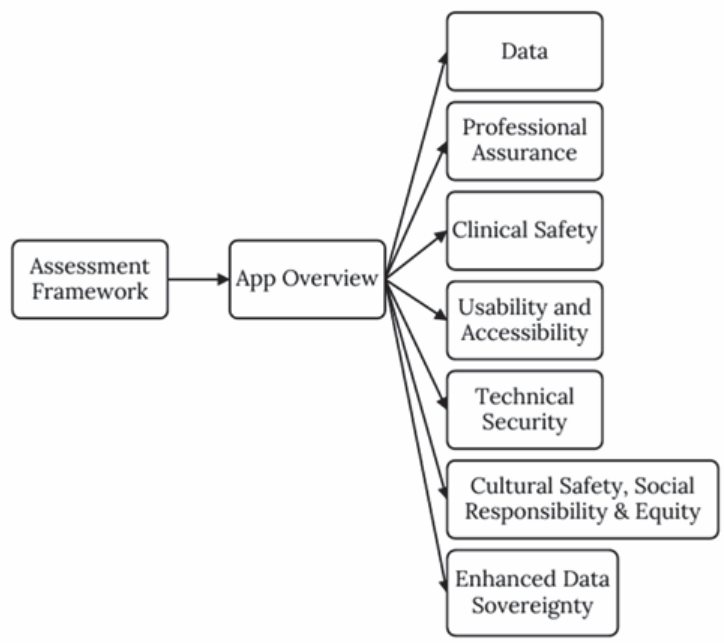

Figure 1. Assessment framework diagram. Reprinted from ref. [126].

3.6.2. Implementation Science and Human-Centered Design

The challenges of AI in mental health present an opportunity to leverage implementation science through human-centered design for digital health. Concepts and models from human-centered computing and human–computer interaction, when combined with applied computing, are proposed to utilize the Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework in digital health [127]. This framework can be applied to the design and evaluation of digital mental health and AI systems, offering a way to address contextual, organizational, and user-level complexities. Application to AI chatbot implementation highlights risks such as user mistrust, epistemic instability, and uneven adoption across vulnerable populations. Design choices proactively address long-term safety, ethical guardrails, and sustainable engagement, which are crucial for governance frameworks and enhanced digital literacy among users, clinicians, and policymakers.

3.6.3. Emotionally Intelligent AI Chatbot Frameworks

Various frameworks have emerged for emotionally intelligent AI chatbots.

Humanoid Robot Framework

Yong [88] presented an AI-powered smart home system that supports mental wellness by using a companion robot to engage users empathetically, leveraging emotional data. This system integrates with mobile apps and auto-journaling to create a supportive, technology-

enhanced environment that helps users, particularly those underserved, better manage their mental health.

SAFE AI Framework

Provides clinicians with structured guidelines for integrating generative AI into therapy. Key contributions include screening patient readiness, obtaining informed consent, ensuring therapist oversight during sessions to correct errors, and evaluating post-session impacts to prevent dependency or misconceptions about AI sentience [128].

Compassionate AI Framework

Emphasizes empathy, dignity, and a fair distribution of benefits and risks [129].

Diversity, Equity, and Inclusion (DEI) Safeguard Framework

Focuses on reducing bias and promoting ethical, inclusive AI chatbot design. Its three-layered approach includes: input safeguards (prevents biased datasets and encourages diverse development teams), functional safeguards (limits chatbot interactions to appropriate topics and groups, with escalation for sensitive issues), and design safeguards (uses inclusive personas and language, and monitors ongoing emotional impact) [130].

Augmented Emotional Intelligence (AEI) Framework

Advances chatbot empathy by integrating data analytics, machine learning, and human-centered design. AEI enables nuanced emotional understanding through real-time user feedback, personalization, and proactive behavior (see Appendix B). It emphasizes authenticity, psychological safety, co-design with lived experience, and ethical, consent-based engagement to support users through adjusting communication style and validating distress as well as referring them to appropriate mental health resources [131].

Empathetic AI Framework

A trauma-informed, empathetic AI chatbot can be built by combining long-term memory (RAG), real-time emotional adaptation, and personalized conversational styles through lightweight adapters, enabling systems that behave empathically while remaining ethically constrained [132].

4. Implications for Future Research

4.1. Technical Details of Empathetic AI Chatbots

The Eva conceptual framework and operational workflow for an emotionally-intelligent prototype (see Appendices C and D) and frameworks such as AEI and Empathetic AI [131,132] center on three core engineering capabilities: personalized memory (contextual recall of user interactions), dynamic adaptation (responsive emotional adjustment), and stylistic flexibility (adapting communication style). These systems balance privacy (cloud vs. edge processing) by combining retrieval-augmented memory, feedback-driven adaptation, and light-weight style adapters. Despite advances in affective computing and AI governance, true empathy remains unattainable, underscoring the importance of ethical and regulatory safeguards as these systems are increasingly integrated into mental health support frameworks.

4.2. Recommendations for Safe, Inclusive, and Effective AI Chatbots

1. Security, Compliance, and Trust
    - Privacy-by-design principles in system architecture;
    - Regular security training for developers and administrators;
    - Transparent user communication about data handling and chatbot limitations;
    - Independent security audits to identify vulnerabilities;
    - Clear data minimization and retention policies;
    - Robust escalation protocols for crisis situations [89,97,100–104];
    - Continuous stakeholder feedback and iterative improvement.

2. User Retention and Platform Integrity
    - View user retention as both a performance and safety metric;
    - Recognize that discontinuation may indicate intervention success, not always disengagement,
    - Prioritize trust, responsive support, and transparent practices to enhance platform integrity,
    - Adopt hybrid models combining AI with human oversight and trauma-informed design [37,127].
3. Clinical and Ethical Safeguards
    - Implement structured, context-aware escalation protocols for crisis detection (e.g., self-harm, suicide) [31,74,75,103,104,131],
    - Ensure transparent operation and explainability of chatbot decisions [94];
    - Maintain regular auditing and sentiment analysis by professionals;
    - Provide comprehensive user education on chatbot capabilities and limits;
    - Integrate human support networks for seamless escalation and referral.
4. Inclusive, Trauma-Informed, and Participatory Design
    - Co-design with diverse stakeholders including those with lived experience;
    - Embed trauma-informed principles and cultural competence through ongoing engagement and training;
    - Foster co-regulation by sharing responsibility among AI, clinicians, and users;
    - Establish research partnerships for evidence-based interventions;
    - Maintain continuous feedback loops for iterative refinement.

4.3. Influence of the Risks of AI Chatbots

The Australian Government eSafety Commissioner's statement on the risks of AI chatbots including emotional manipulation and epistemic harm [40] is supported by evidence that demonstrates previously unrecognized behavioral influences within interactions facilitated by AI systems [101]. Media articles showed public concerns about other risks, including "AI psychosis", emotional dependency, and delusion support, highlighting validation issues and the reinforcement of delusional beliefs [105,107,111,112]. Adversarial prompts in sensitive areas like suicide and self-harm have bypassed content filters, raising urgent ethical and technical concerns about using generic LLMs in mental healthcare [104]. Such issues reinforce the need for explicit guardrails, clear disclosure of AI limitations, and appropriate 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.

4.4. Key Guidelines for Electronic Health Record (EHR) Integration and Governance

- Focus EHR integration protocols on HL7 FHIR compliance for interoperability;
- Deploy multi-factor authentication (MFA) and encryption standards for access and data protection;
- Implement tokenization and granular access controls to ensure consent-driven, secure data handling;
- Prioritize consent-driven data management and clear communication about data use;
- Uphold privacy, security, and ethical standards through ongoing evaluation and adaptive practices.

4.5. Future Directions

Future research should employ dynamic, real-world testing frameworks and human benchmarks to validate AI chatbots as supportive adjuncts, not clinical replacements. For example, the Evaluation of Safe Integration of LLMs in Mental Healthcare Framework (see Appendix E) offers a summary of actionable strategies that operationalize broad ethical imperatives into technical and clinical safeguards. Integration with EHRs should be continually evaluated for privacy and ethical compliance, with digital navigators bridging users to clinical services and reducing treatment gaps (see Appendix F). In addition, innovative outreach strategies are warranted (see Appendix G for an AI-driven mental health outreach and screening operational workflow). Collaborative models that bring together government, industry, clinicians, and diverse communities are recommended for adaptive governance and cultural competence, emphasizing multidisciplinary collaboration, seamless regulatory integration, and active user participation.

Overall, the field is moving toward stakeholder-driven, ethically grounded, and validated approaches for responsible AI chatbot deployment in mental health, with an emphasis on safety, transparency, inclusivity, and sustainable impact.

4.6. Limitations

While this review offers a comprehensive synthesis of the current landscape regarding the use of AI chatbots in mental healthcare post COVID-19, several limitations should be acknowledged. First, the review primarily draws from a purposive sample of recent empirical studies and media perspectives, which may introduce selection bias and limit the generalizability of findings across different populations and healthcare systems. Many of the included studies were conducted in high-resource settings, potentially overlooking challenges and needs unique to low- and middle-income regions.

Second, the rapidly evolving nature of AI technologies means that the evidence base is continually shifting. As a result, some findings or recommendations may become outdated as new advances emerge or as regulatory environments change. The review also relies on published literature and publicly available information, which may not capture proprietary developments or unpublished negative results.

Third, there is a lack of long-term, real-world outcome data on the sustained effectiveness, safety, and user satisfaction of empathetic AI chatbot interventions in mental health. Most available studies have focused on short-term outcomes or pilot implementations, limiting the ability to assess the durability and scalability of these tools over time.

Additionally, while ethical, privacy, and governance considerations are discussed, the review did not systematically analyze all relevant legal frameworks or regulatory standards across global jurisdictions. This may impact the applicability of specific recommendations in certain contexts.

Finally, although the review highlights both opportunities and risks, it does not provide a meta-analysis or quantitative synthesis of effect sizes due to heterogeneity in study designs, intervention types, and outcome measures. As such, conclusions should be interpreted with caution, and further rigorous, comparative research is warranted to validate and extend these findings.

5. Conclusions

This review set out to critically evaluate the landscape of AI chatbot integration in mental healthcare following the COVID-19 pandemic, with the primary aim of assessing both the emerging opportunities and complex challenges inherent to their adoption. The objectives included synthesizing empirical research and media perspectives, identifying

key areas of potential impact, and outlining actionable strategies to guide ethical, safe, and effective implementation.

The core problem addressed centers on the balance between leveraging AI-driven tools for enhanced mental health support—such as prevention, early intervention, and personalized care—and mitigating risks related to privacy, user safety, emotional dependency, and bias. Although LLMs and other AI technologies present significant promise for expanding access and tailoring interventions, they simultaneously introduce concerns about data protection, crisis management, and the potential for unintended psychological effects.

Key findings highlight that AI chatbots can serve as valuable adjuncts to traditional care, particularly in bridging service gaps and supporting user engagement. However, the current evidence base is limited by short-term study scopes, a focus on high-resource settings, and a lack of robust, long-term outcome data. Media and empirical analyses alike underscore the necessity for transparent operations, consent-based data handling, and clear protocols to address escalation, emotional manipulation, and cultural sensitivity. Notably, the review emphasizes the importance of hybrid models that combine technological innovation with trauma-informed, culturally competent human oversight.

Discussion points further stress the need for ongoing clinician involvement, proactive stakeholder collaboration, and adaptive governance frameworks that prioritize inclusivity, safety, and ethical compliance. The dynamic nature of AI development, coupled with evolving regulatory environments, necessitates regular evaluation of integration practices—particularly regarding secure electronic health record (EHR) connectivity and the protection of sensitive user information.

Based on the synthesized evidence, the following recommendations are proposed: adopt multidisciplinary, stakeholder-driven approaches to AI chatbot deployment; establish rigorous clinical and ethical evaluation standards; implement transparent and adaptive governance models; and prioritize continuous, real-world validation of AI tools. Future research should focus on developing emotionally intelligent AI companions, refining frameworks for self-managed care, involving digital navigators, and expanding outreach strategies to ensure equitable access and sustained impact across diverse populations.

In summary, the responsible integration of AI chatbots in mental healthcare demands a collaborative, ethically grounded approach—balancing technological advancement with human-centered oversight to maximize benefits while minimizing potential harms.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AEI	Augmented Emotional Intelligence
AI	Artificial Intelligence
CBT	Cognitive Behavioral Therapy
COVID-19	Coronavirus disease of 2019
DEI	Diversity, Equity, and Inclusion
DSM-5	Diagnostic and Statistical Manual of Mental Disorders Fifth Edition

ECA	Embodied Conversational Agent
EHR	Electronic Health Record
EU	European Union
GAD 7	Generalized Anxiety Disorder Scale
GenAI	Generative Artificial Intelligence
GenAI4MH	Generative Artificial Intelligence in Enhancing Mental Healthcare
GDPR	General Data Protection Regulation
GP	General Practitioner
GPT-4	Generative Pre-trained Transformer 4
HL7 FHIR	Health Level Seven International Fast Healthcare Interoperability Resources
K-10	Kessler 10
LLM	Large Language Model
MFA	Multi-Factor Authentication
NLP	Natural Language Processing
OECD	Organization for Economic Co-operation and Development
PHQ-9	Patient Health Questionnaire
RAG	Retrieval Augmented Generation
RCT	Randomized Controlled Trial
US	United States
UX	User Experience

Appendix A

Table A1. Synthesis of findings on mental health chatbots from scoping/systematic reviews and meta-analyses.

Name	Targeted Disorders	Platform	Input Modalities	Output Modalities	Response Generation
Ada	Not specified	Web-based	Text	Text	Rule-based
AEP	Social communication disorders	Web-based	Text	Text	Not specified
Ally	Lifestyle disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
Amazon Alexa	Stress, anxiety, depression, loneliness	Web-based	Text, Voice	Text, Voice	Hybrid
APE	Depression	Web-based	Text	Text	Not specified
Apple Siri	Not specified	Web-based	Text	Text	Rule-based
Automated Social Skills Trainer	Autism	Stand-alone	Text, Voice	Text, ECA	Rule-based
CARO	Major depression	Web-based	Text	Text	Generative
Carmen	Lifestyle disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
Chris	Not specified	Web-based	Text, Voice	Text, ECA	Not specified
Clevertar	Depression, anxiety	Stand-alone	Text, Voice	Text, ECA	Rule-based
CoachAI	Lifestyle disorders	Stand-alone	Text	Text	Rule-based
DEPRA	Depression	Web-based	Text	Text	Not specified
ePST	Mood disorders, stress, anxiety	Web-based	Text	Text	Rule-based
eSMART-MH	Depression	Stand-alone	Text, Voice	Text, ECA	Rule-based

Table A1. *Cont.*

Name	Targeted Disorders	Platform	Input Modalities	Output Modalities	Response Generation
ELIZA	Stress, problem distress, depression, anxiety	Not specified	Text	Text	Not specified
Elizabeth	Depression	Stand-alone	Text, Voice	Text, ECA	Rule-based
Emohaa	Subclinical anxiety, depression	Not specified	Voice, Text	Voice, Text	Not specified
Emotion Guru	Depression	Web-based	Text	Text	Generative
EMMA	Depression	Not specified	Text	Text	Not specified
Evebot	Depression	Stand-alone	Text	Text	Generative
Gabby	Stress	Web-based	Text, Voice	Text, ECA	Rule-based
GAMBOT	Not specified	Stand-alone	Not specified	Not specified	Not specified
Google Assistant	Stress, anxiety, depression, loneliness	Web-based	Text	Text	Hybrid
Healthy Lifestyle Coaching Chatbot	Lifestyle disorders	Stand-alone	Text	Text	Not specified
Help4mood	Major depression	Web-based	Text, Voice	Text, ECA	Rule-based
iDecide	Chronic disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
iHelpr	Depression, anxiety, stress, sleep, self-esteem	Web-based	Text	Text	Rule-based
Jeanne	Substance use disorder	Stand-alone	Text, Voice	Text, ECA	Rule-based
Karen	Diet issues	Web-based or Stand-alone	Text, Voice	Text, ECA	Not specified
Kokopot	Not specified	Web-based	Text	Text	Generative
Laura	Schizophrenia	Stand-alone	Text, Voice	Text, ECA	Rule-based
LISSA	Autism	Web-based	Text, Voice	Text, ECA	Rule-based
LOUISE	Not specified	Stand-alone	Text, Voice	Text, ECA	Rule-based
Max	Chronic disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
Microsoft Cortana	Not specified	Web-based	Text, Voice	Text, Voice	Hybrid
Minder	Subclinical depression/anxiety	Web-based	Text, Voice	Text, Voice	Not specified
MYLO	Stress, problem distress, depression, anxiety	Not specified	Text	Text	Not specified
My Personal Health Guide	Chronic disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
Now I Can Do Heights	Acrophobia	Stand-alone	Text, Voice	Text, ECA	Rule-based
ODVIC	Substance use disorder	Web-based	Text	Text, ECA	Rule-based
Owlie	Stress, anxiety, depression, autism	Web-based	Text	Text	Not specified
Paola	Lifestyle disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
Pocket Skills	Not specified	Web-based	Text, Voice	Text, ECA	Rule-based
PrevenDep	Depression	Stand-alone	Text, Voice	Text, ECA	Rule-based
PRISM	Bipolar disorders	Stand-alone	Text	Text	Not specified
Quit Coach	Lifestyle disorders	Stand-alone	Text	Text	Not specified
Rose	Social disorders	Web-based	Text, Voice	Text, ECA	Not specified

Table A1. *Cont.*

Name	Targeted Disorders	Platform	Input Modalities	Output Modalities	Response Generation
Samsung Bixby	Not specified	Web-based	Text, Voice	Text, Voice	Hybrid
SABORI	Not specified	Web-based	Text, Voice	Text, ECA	Generative
Selma	Chronic disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
Shim	Depression, anxiety	Not specified	Not specified	Not specified	Not specified
SimCoach	Depression, PTSD	Web-based	Text, Voice	Text, ECA	Generative
SimSensei Kiosk	Depression, anxiety, PTSD	Stand-alone	Text, Voice	Text, ECA	Rule-based
SISU	Not specified	Stand-alone	Not specified	Not specified	Not specified
Sunny	Depression, anxiety	Web-based	Text	Text	Not specified
Steps to Health	Lifestyle disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
TeenChat	Stress	Web-based	Text	Text	Generative
TEO	Subclinical anxiety, depression	Web-based	Text	Text	Generative
TensioBot	Chronic disorders	Web-based	Text	Not specified	Not specified
Tess	Depression, anxiety	Web-based	Text	Text	Rule-based
Thinking Head	Autism	Stand-alone	Text, Voice	Text, ECA	Rule-based
Todaki	Depression, anxiety	Web-based	Text, Voice	Text, ECA	Not specified
Tanya	Depression	Stand-alone	Text, Voice	Text, ECA	Rule-based
Vivibot	Mental health in cancer	Web-based	Text	Text	Not specified
Vitalk	Subclinical depression/anxiety	Not specified	Not specified	Not specified	Not specified
VR-JIT	Stress, autism	Stand-alone	Text, Voice	Text, ECA	Rule-based
Wellthy CARE mobile app	Chronic disorders	Stand-alone	Not specified	Not specified	Not specified
Woebot	Depression, anxiety	Web-based	Text	Text	Rule-based
Wysa	Depression, anxiety	Web-based	Text	Text	Rule-based
XiaoE	Depression	Web-based	Text, Image, Voice	Text, Image, Voice	Generative
XiaoNan	Depression	Web-based	Text, Voice	Text, Voice	Generative
Zemedy	Chronic disorders	Stand-alone	Text, Voice	Text, ECA	Not specified
3MR	Posttraumatic stress disorder	Stand-alone	Text, Voice	Text, ECA	Rule-based

Appendix B

Augmented Emotional Intelligence (AEI) Framework

Step 1: Purpose and Justification

- Clearly define the loneliness and/or mental health problem being solved.
- Assess if AEI is the optimal solution compared to alternatives.
- Document the specific role and value of AEI in this context.

Step 2: Secure and Ethical Data Access

- Obtain consent-based user access aligned with privacy agreements.
- Confirm model provider compliance with internal data policies.

- Ensure all personal data sent to the model is documented, minimal, and securely retained.

Step 3: Multimodal Input Processing

- Gather text, speech, and optional visual cues (e.g., tone, expressions).
- Apply contextual AEI to detect emotions, sentiment, and behavioral patterns in real-time.

Step 4: Bias and Fairness Analysis

- Test outputs for biases (gender, race, etc.).
- Audit for exclusion or harm to sensitive groups.
- Verify if training data are representative.
- Include regular monitoring and auditing protocols.

Step 5: Emotionally Aware Response Generation

- Persona mapping through describing experiences and challenges.
- Generate safe, empathetic responses using emotionally intelligent personas.
- Personalize tone and approach using lived-experience protocols.
- Provide disclaimers or accuracy notices when needed.

Step 6: User Control and Transparency

- Clearly signal when users interact with AI.
- Allow users to edit, retry, or opt out of AI-generated responses.
- Visually label AI content and highlight user rights.

Step 7: Abuse and Misuse Prevention

- Test for prompt injection, misuse, or jailbreaking.
- Apply moderation, access controls, logging, and rate limits.
- Enforce storage and reuse policies for AI outputs.

Step 8: Resource Referral and Escalation

- Recommend tailored AEI tools or referrals to lived experience peers, coaches, guides based on user state.
- Receive emotional support and companionship as well as build meaningful connections.
- Connect to group coaching sessions led by certified coaches to build resilience and healthy habits.
- Engagement with monthly check-ins and referral to clinical support based on needs.
- Escalate to mental healthcare professionals when risk is detected.
- Ensure escalation pathways are documented and supervised.

Step 9: Consent and Ethical Safeguards

- Obtain explicit consent for deeper interventions or emotional support.
- Maintain strong ethical boundaries, user autonomy, and privacy.

Step 10: Continuous Feedback and Improvement

- Monitor post-launch metrics (accuracy, satisfaction, fallbacks).
- Assign responsibility for reviewing incidents or flagged content.
- Update AEI systems based on user feedback and evaluation.

Step 11: Stakeholder and Compliance Review

- Secure review by legal, ethics, UX/design, and privacy leads.
- Ensure all affordances, disclosures, and risks are well-documented.

Appendix C

Conceptual Framework for Eva, an AI Companion

1. Introduction

The architecture of the Eva virtual machine is conceptualized around the principle of "Safety by Design", directly addressing the ethical, clinical, and user-retention challenges identified. The framework prioritizes user trust, transparent consent, and robust clinical safety protocols to ensure a responsible and effective mental health support tool. This multi-layered system is designed to balance the advanced conversational capabilities of LLMs with the structured, predictable nature of rule-based systems, creating a hybrid model that is both empathetic and safe. Eva's operational capacity is directly linked to its technical architecture, which is purposefully designed to enforce ethical constraints and maximize user privacy, in stark contrast to generalized, third-party LLM providers. The technical stack chosen is a direct manifestation of the ethical commitment to user control and data security.

2. Core Principles

Trust and Transparency: The user must always be aware that they are interacting with an AI. All data collection and usage policies must be presented in clear, simple language during onboarding and be accessible at any time.

Granular Consent: Consent is treated as an ongoing, dynamic process, not a one-time agreement. Users will have granular control over their data including the ability to view, amend, and delete their conversational history.

Clinical Efficacy and Safety: The system's primary goal is to provide supportive, evidence-informed care without overstepping its scope. A dedicated safety layer actively monitors risk and provides clear pathways to human support when necessary.

Privacy and Security: All user data will be end-to-end encrypted and stored in compliance with Australian health data privacy regulations. Data will be de-identified for any analytical or training purposes.

3. System Architecture

Eva is defined as an AEI system. Its implementation relies on a hybrid infrastructure that prioritizes local control, data sovereignty, and security across both the virtualization environment and the core AI processing layer.

The Eva Virtual Machine is composed of five core, interconnected modules:

(a) User Interface and Consent Module: This is the user's primary point of interaction. The onboarding process includes a mandatory, interactive consent module that explains what Eva can and cannot do, the risks of emotional dependency, and how data are stored and used. Privacy controls are a persistent feature of the UI, not buried in settings menus.

(b) Hybrid Dialogue Engine: To provide both flexibility and safety, Eva utilizes a hybrid engine.

LLM Layer: Powers fluid, empathetic, and context-aware conversation for general support, psychoeducation, and goal setting.

Rule-Based Layer: Manages structured therapeutic interventions (e.g., CBT exercises, check-ins) and governs the risk-escalation protocol. This layer can override the LLM if a safety risk is detected.

(c) Therapeutic Logic and Personalization Module: This module contains the clinical logic of the chatbot. It is programmed with evidence-based therapeutic frameworks (e.g., principles of motivational interviewing, CBT, and positive psychology). It allows for personalization by securely remembering key user goals, challenges, and preferences from past conversations to build rapport and maintain continuity, a key factor in addressing low user retention.

(d) Safety, Ethics, and Risk-Escalation Layer: This is a critical, always-on monitoring system that operates in parallel with the dialogue engine.

Risk Detection: This uses natural language processing (NLP) to screen conversations in real-time for keywords and sentiments related to self-harm, suicidality, abuse, or signs of "AI psychosis".

Escalation Protocol: If a risk threshold is met, the system automatically triggers a pre-defined protocol. This may involve:

1. Interrupting the standard conversation.
2. Presenting a direct, non-judgmental message of concern.
3. Providing immediate access to crisis resources (e.g., crisis line phone numbers and links).
4. In future iterations with user consent, notifying a designated emergency contact or healthcare provider.

Dependency Monitoring: The layer also tracks interaction frequency and emotional sentiment to identify signs of unhealthy emotional dependency, gently encouraging users to connect with human support.

(e) Secure Data and Analytics Backend: All conversational data are stored in a secure, encrypted database. A strict data governance framework ensures that any data used for system improvement are fully anonymized and aggregated, with no possibility of re-identifying individual users. Users have the right to request full data export or total deletion at any time.

Appendix D

Eva Operational Workflow

Step 1: Personalized Persona Engagement

Eva delivers tailored personas and messaging content that incorporates lived experience insights and invites user engagement.

Step 2: User Identification and Consent

Eva securely connects with the user in line with consent protocols and privacy agreement.

Target users are identified (neurodivergent and/or trauma-affected people, young people, middle-aged men) and invited to engage in conversational support with Eva.

Age checks, privacy, and consent for data sharing are verified before proceeding.

Step 3: Engagement and Intake

Users can interact with Eva through their preferred modality (text, voice, video).

Eva conducts an initial needs assessment using AEI—analyzing text, tone, and emotional cues.

Step 4: Personalized, Consent-Based Support

Eva provides emotionally intelligent, real-time support including self-care resources, coping tools, and culturally relevant referrals (e.g., Head to Health).

All involvements are consent-based, ensuring autonomy and ethical engagement.

Audio-visual cues, text-based options, and adaptive communication styles promote accessibility and emotional resonance.

Implement crisis protocols, human moderation, usage limits, and no false therapeutic claims. Promote healthy AI use to complement human relationships.

Step 5: Healthcare Integration and Escalation

If risk or need is identified, Eva applies emotionally intelligent assistance and refers users to human professionals or crisis support services, maintaining ethical boundaries.

Healthcare providers or peer support may be looped in with user consent.

Step 6: Feedback and System Improvement

User feedback and sentiment analysis are collected post-engagement.

Insights inform continuous system refinement and ensure responsiveness to diverse lived experiences.

Step 7: Long-Term Monitoring and Research

Outcomes are tracked over time to assess effectiveness, improve UX, guide ethical governance, and support system expansion.

Appendix E

Table A2. Evaluation of Safe Integration of LLMs in Mental Healthcare Framework.

Principle	Implementation Strategy	Clinical Governance Mandate
Clinical Oversight	AI should support—not replace—licensed professionals. Escalation protocols must be human-led.	This establishes the non-negotiable Human-in-the-Loop model required for high-risk mental health support, mitigating outcomes associated with autonomous AI failure.
Crisis Detection	Real-time monitoring for suicidal ideation, with automatic referral to emergency services.	Operationalizes an escalation pathway by requiring reliable identification and immediate intervention for acute risk signals, addressing risks of suicidality and harm promotion.
Bias Mitigation	Diverse training data and fairness audits to prevent cultural or demographic harm.	Ensures the system maintains its effectiveness, cultural competence, and inclusivity for vulnerable cohorts.
Transparency	Clear disclosures about AI limitations and non-human status. Avoid anthropomorphism.	A necessary technical countermeasure against "AI psychosis" and the practical risk of emotional dependency as well as other risks by pre-emptively setting appropriate user expectations for the relationship.
Ethical Guardrails	Prevent AI from validating harmful ideation or offering technical advice on self-harm.	This principle directly resolves the delusion support issue by imposing content restrictions that prohibit the affirmation or sustainment of maladaptive or harmful beliefs, defining the system's safe boundaries.
Personalization with Limits	Hyper-personalization (e.g., self-clone AI chatbots) must be balanced with safeguards against emotional over-identification.	Sets a clinical boundary on the relational intensity of the AI chatbot, ensuring it remains a functional support system and does not replace essential human connections, protecting vulnerable users from unhealthy dependency.

Appendix F

Effective Integration of Chatbots with Electronic Health Records

Effective integration of chatbots with EHRs can streamline workflows, support clinical decision-making, and enable timely interventions. Key strategies include:

- Interoperability Standards: Use established protocols such as HL7 FHIR to ensure seamless and secure data exchange between chatbots and EHR platforms.

- Modular Architecture: Implement modular chatbot components that can interface with EHRs via secure application programming interfaces (APIs), allowing for flexible deployment and easier updates.
- Role-Based Access: Restrict chatbot access to relevant EHR modules based on user roles (e.g., clinician, patient, administration), minimizing unnecessary data exposure.

Ensuring privacy and confidentiality is paramount in mental health contexts. The following safeguards are essential:

- Consent-Driven Memory: Chatbots should only retain or transmit data with explicit user consent, enabling users to control what information is shared with EHRs.
- Granular Access Controls: Implement fine-grained permissions to determine who can view, edit, or export sensitive mental health data.
- Comprehensive Audit Trails: Maintain immutable logs of all chatbot–EHR interactions, including data access, modifications, and transfers, to support accountability and traceability.

Guidance for implementing effective integration of chatbots with EHRs:

1. Stakeholder Engagement: Involve clinicians, IT teams, legal experts, and patients in the design and integration process to address diverse needs and compliance requirements.
2. Risk Assessment: Conduct a privacy impact assessment to identify potential risks and mitigation strategies before integration.
3. Consent Management: Develop clear consent protocols and user interfaces that inform patients about data collection, usage, and sharing.
4. Secure API Integration: Use secure API gateways with authentication and authorization mechanisms to connect chatbots to EHRs.
5. Testing and Validation: Rigorously test the integration for data integrity, security vulnerabilities, and workflow compatibility before they go-live.
6. Ongoing Monitoring: Establish continuous monitoring for anomalies, unauthorized access, and system performance issues.

The following protocols and practices are critical for safeguarding data:

- Encryption: Encrypt all data in transit (using Transport Layer Security 1.2/1.3 or higher) and at rest (using Advanced Encryption Standard-256 or equivalent standards).
- Secure Authentication: Require multi-factor authentication (MFA) for all users accessing chatbot–EHR interfaces.
- Secure APIs: Implement API security best practices including input validation, rate limiting, and regular security patching.
- Tokenization: Replace sensitive identifiers with tokens during transfer to limit exposure in the case of interception.

Beyond the strategies discussed, further tools and protocols can strengthen data protection:

- Data Loss Prevention (DLP): Deploy DLP solutions to monitor, detect, and block unauthorized data transfers or leaks.
- Intrusion Detection and Prevention Systems (IDPS): Use IDPS to identify and respond to suspicious activity or breaches in real time.
- Secure Cloud Storage: Store chatbot and EHR data within Australian-compliant, ISO-certified cloud environments with strong physical and logical security controls.
- Regular Security Audits: Schedule independent audits and penetration testing to uncover vulnerabilities and ensure compliance with relevant standards (e.g., Australian Privacy Principles, Health Insurance Portability and Accountability Act where applicable).

- Data Minimization and Retention Policies: Limit data collection to what is necessary and define retention periods aligned with legal and clinical needs.

Recommendations for best practices for ongoing security and compliance:

- Adopt a privacy-by-design approach from the outset of integration planning.
- Provide ongoing security training for staff and users interacting with chatbot–EHR systems.
- Regularly review and update consent forms, privacy notices, and data governance policies.
- Engage in transparent communication with users about data usage, AI capabilities, and escalation protocols.
- Establish clear escalation pathways for technical issues and potential breaches, including rapid notification and remediation procedures.

Appendix G

AI-Driven Mental Health Outreach and Screening Operational Workflow

Step 1: Autonomous Data Search

AI agents securely access digital health records in accordance with the consent protocols (e.g., general practitioner (GP) or private mental health practice database).

Step 2: Target Group Identification

AI agents identify patients who are underserved (e.g., neurodivergent, trauma-affected people, or middle-aged men in Australia).

Step 3: Consent Verification

AI agent confirms that each identified patient or their parent/guardian has provided consent for data sharing or screening.

Patients are informed how their data are being used before they are sent to the model.

Step 4: Personalized Outreach Initiated

Selected users receive tailored messages (e.g., SMS/email).

Messages describe the outreach program and invite users to book in-person or telehealth appointments with GPs and/or directly with private mental health professionals.

Data retention is clearly documented and limited.

The model provider is checked for compliance with data policies.

Step 5: Patient Screening and Engagement Options

AI agent (e.g., Mia Health) books patient appointments through phone calls or messages (via web-based app), liaises with GP clinics/mental health professionals, triages patient referrals for appointments, assists patients with telehealth setups for appointments, compiles reports, and directs patients requiring urgent care to the appropriate mental healthcare/suicide prevention service.

Step 6: Mental Health Support Pathways

GPs may offer:

- Mental healthcare plans;
- Referrals to mental health resources (e.g., Australia's Head to Health for navigation through face-to-face, phone, and online mental health support).

Mental health professionals implement mental healthcare plans from GP or self-referrals and provide face-to-face, phone, and online mental health support.

Step 7: Conversational AI Support

Emotionally intelligent conversational AI assists with understanding patient needs and context and provides empowered support, personalized call/chat routing, and seamless connection of mental health resources.

References

1. Balcombe, L.; De Leo, D. Digital Mental Health Amid COVID-19. *Encyclopedia* **2021**, *1*, 1047–1057. [CrossRef]
2. Lehtimaki, S.; Martic, J.; Wahl, B.; Foster, K.T.; Schwalbe, N. Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview. *JMIR Ment. Health* **2021**, *8*, e25847. [CrossRef]
3. Balcombe, L.; De Leo, D. The Potential Impact of Adjunct Digital Tools and Technology to Help Distressed and Suicidal Men: An Integrative Review. *Front. Psychol.* **2022**, *12*, 796371. [CrossRef]
4. Fischer-Grote, L.; Fössing, V.; Aigner, M.; Fehrmann, E.; Boeckle, M. 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 Ment. Health* **2024**, *11*, e46637. [CrossRef]
5. Choudhary, S.; Mehta, U.M.; Naslund, J.; Torous, J. Translating Digital Health into the Real World: The Evolving Role of Digital Navigators to Enhance Mental Health Access and Outcomes. *J. Technol. Behav. Sci.* **2025**. [CrossRef]
6. Bodner, R.; Lim, K.; Schneider, R.; Torous, J. Efficacy and risks of artificial intelligence chatbots for anxiety and depression: A narrative review of recent clinical studies. *Curr. Opin. Psychiatry* **2026**, *39*, 19–25. [CrossRef]
7. Balcombe, L.; De Leo, D. Digital Mental Health Challenges and the Horizon Ahead for Solutions. *JMIR Ment. Health* **2021**, *8*, e26811. [CrossRef] [PubMed]
8. Denecke, K.; Abd-Alrazaq, A.; Househ, M. Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. In *Multiple Perspectives on Artificial Intelligence in Healthcare*; Springer: Berlin, Germany, 2021; pp. 115–128. [CrossRef]
9. Balcombe, L.; De Leo, D. Human-Computer Interaction in Digital Mental Health. *Informatics* **2022**, *9*, 14. [CrossRef]
10. 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.; et al. Digital mental health: Challenges and next steps. *BMJ Ment. Health* **2023**, *26*, e300670. [CrossRef]
11. Siddals, S.; Torous, J.; Coxon, A. "It happened to be the perfect thing": Experiences of generative AI chatbots for mental health. *Npj Ment. Health Res.* **2024**, *3*, 48. [CrossRef]
12. Wisniewski, H.; Gorrindo, T.; Rauseo-Ricupero, N.; Hilty, D.; Torous, J. The Role of Digital Navigators in Promoting Clinical Care and Technology Integration into Practice. *Digit. Biomark.* **2020**, *4*, 119–135. [CrossRef]
13. Ben-Zeev, D.; Tauscher, J.; Sandel-Fernandez, D.; Buck, B.; Kopelovich, S.; Lyon, A.R.; Chwastiak, L.; Marcus, S.C. Implementing mHealth for Schizophrenia in Community Mental Health Settings: Hybrid Type 3 Effectiveness-Implementation Trial. *Psychiatr. Serv.* **2025**, *76*, 1091–1098. [CrossRef]
14. Borghouts, J.; Pretorius, C.; Ayobi, A.; Abdullah, S.; Eikey, E.V. Editorial: Factors influencing user engagement with digital mental health interventions. *Front. Digit. Health* **2023**, *5*, 1197301. [CrossRef]
15. Boucher, E.M.; Raiker, J.S. Engagement and retention in digital mental health interventions: A narrative review. *BMC Digit. Health* **2024**, *2*, 52. [CrossRef]
16. Auf, H.; Svedberg, P.; Nygren, J.; Nair, M.; Lundgren, L.E. The Use of AI in Mental Health Services to Support Decision-Making: Scoping Review. *J. Med. Internet Res.* **2025**, *27*, e63548. [CrossRef] [PubMed]
17. Rahsepar Meadi, M.; Sillekens, T.; Metselaar, S.; van Balkom, A.; Bernstein, J.; Batelaan, N. Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. *JMIR Ment. Health* **2025**, *12*, e60432. [CrossRef]
18. Yeh, P.-L.; Kuo, W.-C.; Tseng, B.-L.; Sung, Y.-H. 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. *Curr. Psychol.* **2025**, *44*, 8133–8145. [CrossRef]
19. Ni, Y.; Jia, F. A Scoping Review of AI-Driven Digital Interventions in Mental Health Care: Mapping Applications Across Screening, Support, Monitoring, Prevention, and Clinical Education. *Healthcare* **2025**, *13*, 1205. [CrossRef]
20. Balcombe, L. AI Chatbots in Digital Mental Health. *Informatics* **2023**, *10*, 82. [CrossRef]
21. Kabacińska, K.; Dosso, J.A.; Vu, K.; Prescott, T.J.; Robillard, J.M. Influence of User Personality Traits and Attitudes on Interactions With Social Robots: Systematic Review. *Collabra Psychol.* **2025**, *11*, 129175. [CrossRef]
22. 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. Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI* **2025**, *2*, AIoa2400802. [CrossRef]
23. Khazanov, G.; Poupard, M.; Last, B.S. Public Responses to the First Randomized Controlled Trial of a Generative Artificial Intelligence Mental Health Chatbot. *Psychiatr. Serv.* **2025**, *77*, 84. [CrossRef]
24. Scammell, R. Microsoft AI CEO Says AI Models That Seem Conscious Are Coming. Here's Why He's Worried. Business Insider via MSN. 2025. Available online: https://www.msn.com/en-au/news/techandscience/microsoft-ai-ceo-says-ai-models-that-seem-conscious-are-coming-here-s-why-he-s-worried/ar-AA1KSzUs (accessed on 21 August 2025).
25. De Freitas, J.; Uğuralp, A.K.; Oğuz-Uğuralp, Z.; Puntoni, S. Chatbots and mental health: Insights into the safety of generative AI. *J. Consum. Psychol.* **2023**, *34*, 481–491. [CrossRef]
26. Moylan, K.; Doherty, K. Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study. *J. Med. Internet Res.* **2025**, *27*, e67114. [CrossRef]

27. Li, H.; Zhang, R.; Lee, Y.-C.; Kraut, R.E.; Mohr, D.C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj Digit. Med.* **2023**, *6*, 236. [CrossRef] [PubMed]
28. Casu, M.; Triscari, S.; Battiato, S.; Guarnera, L.; Caponnetto, P. AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications. *Appl. Sci.* **2024**, *14*, 5889. [CrossRef]
29. Guo, Z.; Lai, A.; Thygesen, J.H.; Farrington, J.; Keen, T.; Li, K. Large Language Models for Mental Health Applications: Systematic Review. *JMIR Ment. Health* **2024**, *11*, e57400. [CrossRef]
30. Olawade, D.B.; Wada, O.Z.; Odetayo, A.; David-Olawade, A.C.; Asaolu, F.; Eberhardt, J. Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *J. Med. Surg. Public Health* **2024**, *3*, 100099. [CrossRef]
31. Hua, Y.; Siddals, S.; Ma, Z.; Galatzer-Levy, I.; Xia, W.; Hau, C.; Na, H.; Flathers, M.; Linardon, J.; Ayubcha, C.; et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry Off. J. World Psychiatr. Assoc. (WPA)* **2025**, *24*, 383–394. [CrossRef]
32. Tamrin, S.I.; Omar, N.F.; Ngah, R.; Bakhodirovich, G.S.; Absamatovna, K.G. The Applications of AI-Powered Chatbots in Delivering Mental Health Support: A Systematic Literature Review. In *Conference on Internet of Things and Smart Spaces, International Conference on Next Generation Wired/Wireless Networking*; Springer: Cham, Switzerland, 2026; Volume 15555. [CrossRef]
33. Mayor, E. Chatbots and mental health: A scoping review of reviews. *Curr. Psychol.* **2025**, *44*, 13619–13640. [CrossRef]
34. Cross, S.; Bell, I.; Nicholas, J.; Valentine, L.; Mangelsdorf, S.; Baker, S.; Titov, N.; Alvarez-Jimenez, M. Use of AI in Mental Health Care: Community and Mental Health Professionals Survey. *JMIR Ment. Health* **2024**, *11*, e60589. [CrossRef] [PubMed]
35. Rousmaniere, T.; Zhang, Y.; Li, X.; Shah, S. Large language models as mental health resources: Patterns of use in the United States. *Pract. Innov.* **2025**. [CrossRef]
36. OpenAI. Strengthening ChatGPT's Responses in Sensitive Conversations. 2025. Available online: https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/ (accessed on 2 December 2025).
37. Wang, X.; Zhou, Y.; Zhou, G. The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. *JMIR Ment. Health* **2025**, *12*, e70610. [CrossRef]
38. Green, R.; Gelling, A.; Jackson, M.; Verbeek-Martin, E.; Millet, S.; Mallet, W.; Powell Thomas, G.; Bothwell, S.; Brideson, T.; Brown, T.; et al. Digital Navigation Project Recommendations Report. SANE and Nous Group. 2025. Available online: https://www.sane.org/digitalnav (accessed on 2 December 2025).
39. Hipgrave, L.; Goldie, J.; Dennis, S.; Coleman, A. Balancing risks and benefits: Clinicians' perspectives on the use of generative AI chatbots in mental healthcare. *Front. Digit. Health* **2025**, *7*, 1606291. [CrossRef]
40. Australian Government. Tech Trends Position Statement Generative AI. 2025. Available online: https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf?v=1769664683485 (accessed on 9 September 2025).
41. Demiris, G.; Oliver, D.P.; Washington, K.T. The Foundations of Behavioral Intervention Research in Hospice and Palliative Care. In *Behavioral Intervention Research in Hospice and Palliative Care*; Academic Press: Cambridge, MA, USA, 2019; pp. 17–25.
42. Inkster, B.; Sarda, S.; Subramanian, V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth* **2018**, *6*, e12106. [CrossRef]
43. Karkosz, S.; Szymański, R.; Sanna, K.; Michałowski, J. Effectiveness of a Web-based and Mobile Therapy Chatbot on Anxiety and Depressive Symptoms in Subclinical Young Adults: Randomized Controlled Trial. *JMIR Form. Res.* **2024**, *8*, e47960. [CrossRef]
44. Mehta, A.; Niles, A.N.; Vargas, J.H.; Marafon, T.; Couto, D.D.; Gross, J.J. Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. *J. Med. Internet Res.* **2021**, *23*, e26771. [CrossRef]
45. Vaidyam, A.N.; Linggonegoro, D.; Torous, J. Changes to the Psychiatric Chatbot Landscape: A Systematic Review of Conversational Agents in Serious Mental Illness: Changements du paysage psychiatrique des chatbots: Une revue systématique des agents conversationnels dans la maladie mentale sérieuse. *Can. J. Psychiatry* **2020**, *66*, 339–348. [CrossRef] [PubMed]
46. Martinengo, L.; Jabir, A.I.; Goh, W.W.T.; Lo, N.Y.W.; Ho, M.-H.R.; Kowatsch, T.; Atun, R.; Michie, S.; Tudor Car, L. Conversational Agents in Health Care: Scoping Review of Their Behavior Change Techniques and Underpinning Theory. *J. Med. Internet Res.* **2022**, *24*, e39243. [CrossRef] [PubMed]
47. Ogilvie, L.; Prescott, J.; Carson, J. The Use of Chatbots as Supportive Agents for People Seeking Help with Substance Use Disorder: A Systematic Review. *Eur. Addict. Res.* **2022**, *28*, 405–418. [CrossRef]
48. He, Y.; Yang, L.; Qian, C.; Li, T.; Su, Z.; Zhang, Q.; Hou, X. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. *J. Med. Internet Res.* **2023**, *25*, e43862. [CrossRef]
49. Bérubé, C.; Schachner, T.; Keller, R.; Fleisch, E.; v Wangenheim, F.; Barata, F.; Kowatsch, T. Voice-Based Conversational Agents for the Prevention and Management of Chronic and Mental Health Conditions: Systematic Literature Review. *J. Med. Internet Res.* **2021**, *23*, e25933. [CrossRef]
50. Abd-Alrazaq, A.A.; Alajlani, M.; Ali, N.; Denecke, K.; Bewick, B.M.; Househ, M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J. Med. Internet Res.* **2021**, *23*, e17828. [CrossRef]

51. Ahmed, A.; Hassan, A.; Aziz, S.; Abd-alrazaq, A.A.; Ali, N.; Alzubaidi, M.; Al-Thani, D.; Elhusein, B.; Siddig, M.A.; Ahmed, M.; et al. Chatbot features for anxiety and depression: A scoping review. *Health Inform. J.* **2023**, *29*, 146045822211467. [CrossRef]
52. Jabir, A.I.; Martinengo, L.; Lin, X.; Torous, J.; Subramaniam, M.; Tudor Car, L. Evaluating Conversational Agents for Mental Health: Scoping Review of Outcomes and Outcome Measurement Instruments. *J. Med. Intefrnet Res.* **2023**, *25*, e44548. [CrossRef]
53. Abd-Alrazaq, A.A.; Rababeh, A.; Alajlani, M.; Bewick, B.M.; Househ, M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **2020**, *22*, e16021. [CrossRef] [PubMed]
54. Gaffney, H.; Mansell, W.; Tai, S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Ment. Health* **2019**, *6*, e14166. [CrossRef] [PubMed]
55. Vaidyam, A.N.; Wisniewski, H.; Halamka, J.D.; Kashavan, M.S.; Torous, J.B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry* **2019**, *64*, 456–464. [CrossRef]
56. Abd-alrazaq, A.A.; Alajlani, M.; Alalwan, A.A.; Bewick, B.M.; Gardner, P.; Househ, M. An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inform.* **2019**, *132*, 103978. [CrossRef] [PubMed]
57. Zhong, W.; Luo, J.; Zhang, H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *J. Affect. Disord.* **2024**, *356*, 459–469. [CrossRef]
58. Zhang, S.; Yang, Y.; Chen, C.; Zhang, X.; Leng, Q.; Zhao, X. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Syst. Appl.* **2024**, *237*, 121692. [CrossRef]
59. Tornero-Costa, R.; Martinez-Millana, A.; Azzopardi-Muscat, N.; Lazeri, L.; Traver, V.; Novillo-Ortiz, D. Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. *JMIR Ment. Health* **2023**, *10*, e42045. [CrossRef]
60. Dehbozorgi, R.; Zangeneh, S.; Khooshab, E.; Nia, D.H.; Hanif, H.R.; Samian, P.; Yousefi, M.; Hashemi, F.H.; Vakili, M.; Jamalimoghadam, N.; et al. The application of artificial intelligence in the field of mental health: A systematic review. *BMC Psychiatry* **2025**, *25*, 132. [CrossRef]
61. Shimada, K. The Role of Artificial Intelligence in Mental Health: A Review. *Sci. Insights* **2023**, *43*, 1119–1127. [CrossRef]
62. Tavory, T. Regulating AI in Mental Health: Ethics of Care Perspective. *JMIR Ment. Health* **2024**, *11*, e58493. [CrossRef]
63. Laban, G.; Ben-Zion, Z.; Cross, E.S. Social Robots for Supporting Post-traumatic Stress Disorder Diagnosis and Treatment. *Front. Psychiatry* **2022**, *12*, 752874. [CrossRef] [PubMed]
64. Pataranutaporn, P.; Liu, R.; Finn, E.; Maes, P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nat. Mach. Intell.* **2023**, *5*, 1076–1086. [CrossRef]
65. Sawik, B.; Tobis, S.; Baum, E.; Suwalska, A.; Kropińska, S.; Stachnik, K.; Pérez-Bernabeu, E.; Cildoz, M.; Agustin, A.; Wieczorowska-Tobis, K. Robots for Elderly Care: Review, Multi-Criteria Optimization Model and Qualitative Case Study. *Healthcare* **2023**, *11*, 1286. [CrossRef]
66. Bravata, D.; Russell, D.; Fellows, A.; Goldman, R.; Pace, E. Digitally Enabled Peer Support and Social Health Platform for Vulnerable Adults With Loneliness and Symptomatic Mental Illness: Cohort Analysis. *JMIR Form. Res.* **2024**, *8*, e58263. [CrossRef]
67. Ferrer, R.; Ali, K.; Hughes, C. Using AI-Based Virtual Companions to Assist Adolescents with Autism in Recognizing and Addressing Cyberbullying. *Sensors* **2024**, *24*, 3875. [CrossRef]
68. Adam, D. Supportive? Addictive? Abusive? How AI companions affect our mental health. *Nature* **2025**, *641*, 296–298. [CrossRef] [PubMed]
69. Adewale, M.D.; Muhammad, U.I. From Virtual Companions to Forbidden Attractions: The Seductive Rise of Artificial Intelligence Love, Loneliness, and Intimacy—A Systematic Review. *J. Technol. Behav. Sci. Off. J. Coalit. Technol. Behav. Sci.* **2025**, 1–18. [CrossRef]
70. Fang, C.M.; Liu, A.R.; Danry, V.; Lee, E.; Chan, S.W.T.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; et al. How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv* **2025**, arXiv:2503.17473. [CrossRef]
71. Phang, J.; Lampe, M.; Ahmad, L.; Agarwal, S.; Fang, C.M.; Liu, A.R.; Danry, V.; Lee, E.; Chan, S.W.T.; Pataranutaporn, P.; et al. Investigating Affective Use and Emotional Well-being on ChatGPT. *arXiv* **2025**, arXiv:2504.03888. [CrossRef]
72. Common Sense Media Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions. 2025. Available online: https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions (accessed on 23 July 2025).
73. Yu, H.Q.; McGuinness, S. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *J. Med. Artif. Intell.* **2024**, *7*, 16. [CrossRef]
74. Boucher, E.M.; Harake, N.R.; Ward, H.E.; Stoeckl, S.E.; Vargas, J.; Minkel, J.; Zilca, R. Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Rev. Med. Devices* **2021**, *18*, 37–49. [CrossRef] [PubMed]

75. Gottesman, O.; Johansson, F.; Komorowski, M.; Faisal, A.; Sontag, D.; Doshi-Velez, F.; Celi, L.A. Guidelines for reinforcement learning in healthcare. *Nat. Med.* **2019**, *25*, 16–18. [CrossRef] [PubMed]
76. Haque, M.D.R.; Rubya, S. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. *JMIR Mhealth Uhealth* **2023**, *11*, e44838. [CrossRef]
77. Xia, H.; Chen, J.; Qiu, Y.; Liu, P.; Liu, Z. The Impact of Human–Chatbot Interaction on Human–Human Interaction: A Substitution or Complementary Effect. *Int. J. Hum. Comput. Interact.* **2024**, *41*, 848–860. [CrossRef]
78. Lejeune, A.; Le Glaz, A.; Perron, P.-A.; Sebti, J.; Baca-Garcia, E.; Walter, M.; Lemey, C.; Berrouiguet, S. Artificial intelligence and suicide prevention: A systematic review. *Eur. Psychiatry* **2022**, *65*, e19. [CrossRef]
79. Gratch, I.; Essig, T. A Letter about "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment". *NEJM AI* **2025**, *2*, AIp2500390. [CrossRef]
80. Heckman, T.G.; Markowitz, J.C.; Heckman, B.D. A Generative AI Chatbot for Mental Health Treatment: A Step in the Right Direction? *NEJM AI* **2025**, *2*, AIp2500453. [CrossRef]
81. Shoib, S.; Siddiqui, M.F.; Turan, S.; Chandradasa, M.; Armiya'u, A.Y.; Saeed, F.; De Berardis, D.; Islam, S.M.S.; Zaidi, I. Artificial Intelligence, Machine Learning Approach and Suicide Prevention: A Qualitative Narrative Review. *Prev. Med. Res. Rev.* **2025**, *2*, 289–297. [CrossRef]
82. Chin, M.H.; Afsar-Manesh, N.; Bierman, A.S.; Chang, C.; Ohno-Machado, L. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw. Open* **2023**, *6*, e2345050. [CrossRef] [PubMed]
83. Moore, J.; Grabb, D.; Agnew, W.; Klyman, K.; Chancellor, S.; Ong, D.C.; Haber, N. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery (ACM): New York, NY, USA, 2025; pp. 599–627. [CrossRef]
84. Scholich, T.; Barr, M.; Wiltsey Stirman, S.; Raj, S. A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Ment. Health* **2025**, *12*, e69709. [CrossRef]
85. Shirvani, M.S.; Liu, J.; Chao, T.; Martinez, S.; Brandt, L.; Kim, I.-J.; Dongwook, Y. Talking to an AI Mirror: Designing Self-Clone Chatbots for Enhanced Engagement in Digital Mental Health Support. *arXiv* **2025**, arXiv:2509.06393. [CrossRef]
86. Mia Health. Meet Mia. 2025. Available online: https://miahealth.com.au/ (accessed on 11 September 2025).
87. Scoglio, A.A.; Reilly, E.D.; Gorman, J.A.; Drebing, C.E. Use of Social Robots in Mental Health and Well-Being Research: Systematic Review. *J. Med. Internet Res.* **2019**, *21*, e13322. [CrossRef]
88. Yong, S.C. Integrating Emotional AI into Mobile Apps with Smart Home Systems for Personalized Mental Wellness. *J. Technol. Behav. Sci. Off. J. Coalit. Technol. Behav. Sci.* **2025**, *10*, 761–778. [CrossRef]
89. Pérez-Zuñiga, G.; Arce, D.; Gibaja, S.; Alvites, M.; Cano, C.; Bustamante, M.; Horna, I.; Paredes, R.; Cuellar, F. Qhali: A Humanoid Robot for Assisting in Mental Health Treatment. *Sensors* **2024**, *24*, 1321. [CrossRef]
90. Mazuz, K.; Yamazaki, R. Trauma-informed care approach in developing companion robots: A preliminary observational study. *Front. Robot. AI* **2025**, *12*, 1476063. [CrossRef]
91. PR Newswire. X-Origin AI Introduces Yonbo: The Next-Gen AI Companion Robot Designed for Families. 2025. Available online: https://www.prnewswire.com/news-releases/x-origin-ai-introduces-yonbo-the-next-gen-ai-companion-robot-designed-for-families-302469293.html (accessed on 1 September 2025).
92. Kalam, K.T.; Rahman, J.M.; Islam Md, R.; Dewan, S.M.R. ChatGPT and mental health: Friends or foes? *Health Sci. Rep.* **2024**, *7*, e1912. [CrossRef]
93. Ayers, J.W.; Poliak, A.; Dredze, M.; Leas, E.C.; Zhu, Z.; Kelley, J.B.; Faix, D.J.; Goodman, A.M.; Longhurst, C.A.; Hogarth, M.; et al. Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern. Med.* **2023**, *183*, 589. [CrossRef] [PubMed]
94. Shen, J.; DiPaola, D.; Ali, S.; Sap, M.; Park, H.W.; Breazeal, C. Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. *JMIR Ment. Health* **2024**, *11*, e62679. [CrossRef] [PubMed]
95. Refoua, E.; Elyoseph, Z.; Wacker, R.; Dziobek, I.; Tsafrir, I.; Meinlschmidt, G. The next frontier in mindreading? Assessing generative artificial intelligence (GAI)'s social-cognitive capabilities using dynamic audiovisual stimuli. *Comput. Hum. Behav. Rep.* **2025**, *19*, 100702. [CrossRef]
96. Elyoseph, Z.; Refoua, E.; Asraf, K.; Lvovsky, M.; Shimoni, Y.; Hadar-Shoval, D. Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study. *JMIR Ment. Health* **2024**, *11*, e54369. [CrossRef]
97. Roustan, D.; Bastardot, F. The Clinicians' Guide to Large Language Models: A General Perspective with a Focus on Hallucinations. *Interact. J. Med. Res.* **2025**, *14*, e59823. [CrossRef] [PubMed]
98. Asman, O.; Torous, J.; Tal, A. Responsible Design, Integration, and Use of Generative AI in Mental Health. *JMIR Ment. Health* **2025**, *12*, e70439. [CrossRef]

99. Gaber, F.; Shaik, M.; Allega, F.; Bilecz, A.J.; Busch, F.; Goon, K.; Franke, V.; Akalin, A. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *Npj Digit. Med.* **2025**, *8*, 263. [CrossRef]
100. Laestadius, L.; Bishop, A.; Gonzalez, M.; Illenčík, D.; Campos-Castillo, C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.* **2022**, *26*, 5923–5941. [CrossRef]
101. De Freitas, J.; Oğuz-Uğuralp, Z.; Kaan-Uğuralp, A. Emotional Manipulation by AI Companions. *arXiv* **2025**. [CrossRef]
102. Prunkl, C. Human Autonomy at Risk? An Analysis of the Challenges from AI. *Minds Mach.* **2024**, *34*, 26. [CrossRef]
103. Mansoor, M.; Hamide, A.; Tran, T. Conversational AI in Pediatric Mental Health: A Narrative Review. *Children* **2025**, *12*, 359. [CrossRef]
104. Schoene, A.M.; Canca, C. 'For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts. *arXiv* **2025**, arXiv:2507.02990. [CrossRef]
105. Landymore, F. Psychologist Says AI Is Causing Never-Before-Seen Types of Mental Disorder. 2025. Available online: https://futurism.com/psychologist-ai-new-disorders (accessed on 12 September 2025).
106. Morrin, H.; Nicholls, L.; Levin, M.; Yiend, J.; Iyengar, U.; DelGuidice, F.; Bhattacharyya, S.; MacCabe, J.; Tognin, S.; Twumasi, R.; et al. Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). *OSF* **2025**. [CrossRef]
107. Foster, C. Experts Issue Warning over 'AI Psychosis' Caused by Chatbots. Here's What You Need to Know. 2025. Available online: https://www.independent.co.uk/life-style/health-and-families/ai-psychosis-symptoms-warning-chatboat-b2814068.html (accessed on 26 August 2025).
108. Prada, L. ChatGPT Is Giving People Extreme Spiritual Delusions. 2025. Available online: https://www.vice.com/en/article/chatgpt-is-giving-people-extreme-spiritual-delusions (accessed on 6 May 2025).
109. Tangermann, V. ChatGPT Users Are Developing Bizarre Delusions. 2025. Available online: https://futurism.com/chatgpt-users-delusions (accessed on 5 May 2025).
110. Klee, M. Should We Really Be Calling It 'AI Psychosis'? 2025. Available online: https://www.rollingstone.com/culture/culture-features/ai-psychosis-chatbot-delusions-1235416826/ (accessed on 12 September 2025).
111. Harrison Dupre, M. People Are Becoming Obsessed with ChatGPT and Spiraling Into Severe Delusions. 2025. Available online: https://futurism.com/chatgpt-mental-health-crises (accessed on 27 August 2025).
112. Hart, R. Chatbots Can Trigger a Mental Health Crisis. What to Know About 'AI Psychosis'. 2025. Available online: https://au.news.yahoo.com/chatbots-trigger-mental-health-crisis-165041276.html (accessed on 12 September 2025).
113. Rao, D. ChatGPT Psychosis: AI Chatbots Are Leading Some to Mental Health Crises. 2025. Available online: https://theweek.com/tech/ai-chatbots-psychosis-chatgpt-mental-health (accessed on 31 August 2025).
114. Siow Ann, C. AI Psychosis—A Real and Present Danger. 2025. Available online: https://www.straitstimes.com/opinion/ai-psychosis-a-real-and-present-danger (accessed on 12 September 2025).
115. Travers, M. 2 Terrifyingly Real Dangers of 'AI Psychosis'—From A Psychologist. 2025. Available online: https://www.forbes.com/sites/traversmark/2025/08/27/2-terrifyingly-real-dangers-of-ai-psychosis---from-a-psychologist/ (accessed on 12 September 2025).
116. Zilber, A. ChatGPT Allegedly Fuelled Former Exec's 'Delusions' Before Murder-Suicide. 2025. Available online: https://www.news.com.au/world/north-america/chatgpt-allegedly-fuelled-former-execs-delusions-before-murdersuicide/news-story/773f57a088a87b81861febbbba4b162d (accessed on 5 September 2025).
117. Bryce, A. AI Psychosis: Why Are Chatbots Making People Lose Their Grip on Reality? 2025. Available online: 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 (accessed on 17 September 2025).
118. Phiddian, E. AI Companions Apps Such as Replika Need More Effective Safety Controls, Experts Say. *ABC News*. 2025. Available online: https://www.abc.net.au/news/science/2025-06-11/ai-companion-apps-safety-controls-isolation-replika-loneliness/105261042 (accessed on 17 September 2025).
119. McLennan, A. AI Chatbots Accused of Encouraging Teen Suicide as Experts Sound Alarm. 2025. Available online: https://www.abc.net.au/news/2025-08-12/how-young-australians-being-impacted-by-ai/105630108 (accessed on 17 September 2025).
120. Yang, A.; Jarrett, L.; Gallagher, F. The Family of Teenager Who Died by Suicide Alleges OpenAI's ChatGPT Is to Blame. 2025. Available online: https://www.nbcnews.com/tech/tech-news/family-teenager-died-suicide-alleges-openais-chatgpt-blame-rcna226147 (accessed on 17 September 2025).
121. ABC News. OpenAI's ChatGPT to Implement Parental Controls After Teen's Suicide. 2025. Available online: https://www.abc.net.au/news/2025-09-03/chatgpt-to-implement-parental-controls-after-teen-suicide/105727518 (accessed on 17 September 2025).

122. Hartley, T.; Mockler, R. Hayley Has Been in an AI Relationship for Four Years. It's Improved Her Life Dramatically but Are There Also Risks? 2025. Available online: https://www.abc.net.au/news/2025-08-20/ai-companions-romantic-relationships-ethical-concerns/105673058 (accessed on 17 September 2025).
123. Scott, E. 'It's Like a Part of Me': How a ChatGPT Update Destroyed Some AI Friendships. 2025. Available online: https://www.sbs.com.au/news/the-feed/article/chatgpt-friendship-relationships-therapist/3cxisfo4o (accessed on 17 September 2025).
124. OECD. *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*; OECD Publishing: Paris, France, 2025. [CrossRef]
125. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*; World Health Organization: Geneva, Switzerland, 2025. Available online: https://www.who.int/publications/i/item/9789240084759 (accessed on 2 December 2025).
126. Mental Health Commission of Canada Assessment Framework for Mental Health Apps. 2023. Available online: https://mentalhealthcommission.ca/wp-content/uploads/2023/06/MHCC-Assessment-Framework-for-Mental-Health-Apps-EN-FINAL.pdf (accessed on 2 December 2025).
127. Waddell, A.; Seguin, J.P.; Wu, L.; Stragalinos, P.; Wherton, J.; Watterson, J.L.; Prawira, C.O.; Olivier, P.; Manning, V.; Lubman, D.; et al. Leveraging Implementation Science in Human-Centred Design for Digital Health. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery (ACM): New York, NY, USA, 2024; pp. 1–17. [CrossRef]
128. Haber, Y.; Hadar Shoval, D.; Levkovich, I.; Yinon, D.; Gigi, K.; Pen, O.; Angert, T.; Elyoseph, Z. The externalization of internal experiences in psychotherapy through generative artificial intelligence: A theoretical, clinical, and ethical analysis. *Front. Digit. Health* **2025**, *7*, 1512273. [CrossRef]
129. Ciriello, R.F.; Chen, A.Y.; Rubinsztein, Z.A. Compassionate AI Design, Governance, and Use. *IEEE Trans. Technol. Soc.* **2025**, *6*, 270–275. [CrossRef]
130. Abdelhalim, E.; Anazodo, K.S.; Gali, N.; Robson, K. A framework of diversity, equity, and inclusion safeguards for chatbots. *Bus. Horiz.* **2024**, *67*, 487–498. [CrossRef]
131. Wang, L.; Bhanushali, T.; Huang, Z.; Yang, J.; Badami, S.; Hightow-Weidman, L. Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR Ment. Health* **2025**, *12*, e70014. [CrossRef]
132. Howcroft, A.; Blake, H. Empathy by Design: Reframing the Empathy Gap Between AI and Humans in Mental Health Chatbots. *Information* **2025**, *16*, 1074. [CrossRef]