Review

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

Luke Balcombe *

Posted Date: 11 December 2025

doi: 10.20944/preprints202512.1012.v1

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

 doi:10.20944/preprints202512.1012.v1

Review

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

Luke Balcombe *

[1] Australian Institute for Suicide Research and Prevention, School of Applied Psychology, Griffith University, Messines Ridge Road, Mount Gravatt, QLD 4122, Australia

* Correspondence: l.balcombe@griffith.edu.au

Abstract

Digital mental health uses technology—like the Internet, smartphones, wearables, and immersive platforms—to improve access to care. While these resources quickly expanded post COVID-19, ongoing issues include low user retention, poor digital literacy, unclear privacy rules, and limited proof of effectiveness and safety. AI chatbots, also known as agents and assistants that act as a therapist or companion, support mental health by delivering counseling and personalized interactions through various apps and devices. AI chatbots may boost social health and lower loneliness, however, they may also increase dependence and affect emotional outcomes. Their use remains largely unregulated, with concerns about privacy, bias, and ethics. Experiences vary; some users report positive results while others doubt their safety and impact, especially in crisis response. There is a need to better protect vulnerable users and engage the underserved, with input from various individuals with lived experience on what feels safe, supportive, or harmful when interacting with AI chatbots. Proper evaluation, standardized training by digital navigators, and ethical/clinical guidelines are crucial for safe, engaging and effective adoption of AI in mental health care and support.

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

1. Introduction

The mental health treatment gap is worsening since the COVID-19 pandemic which triggered a digital transformation of mental health care 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. 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 regards 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. Consequently, there are renewed calls for trained Digital Navigators to assist in the safe integration of technology into mental health care settings, driving engagement, and supporting both the needs of the clinician and the patient [5,12]. A recent trial found changes in the level of support provided by Digital Navigators directly affect how effective schizophrenia apps are, meaning standardized training is essential to reliably evaluate these tools [13].

 doi:10.20944/preprints202512.1012.v1

GenAI-powered platforms, especially those using advanced Large Language Models (LLMs), are increasingly sought out for consulting about mental health care. 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 showed they generate coherent text, maintain conversational context, and perform advisory or counseling tasks, making them suitable for various mental health applications [19]. Nonetheless, the integration of AI chatbots into clinical practice is understudied in terms of effectiveness and safety, and the ability to provide nuanced, meaningful support.

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 health care 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.

There is a lack of clinical evidence supporting AI-based therapy because of limited conclusions about 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 yet 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 have been cases where AI chatbots appeared to exhibit consciousness—a phenomenon referred to as "Seemingly Conscious AI"—which, despite a lack of scientific evidence for AI consciousness, has prompted debate about ethical design, transparency, and the need for human oversight [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]. It is unclear how many people are using AI chatbots for mental health support globally, although findings in Australia with community members show it to be 28% and 43% for mental health professionals [34]. A 2025 survey of US residents with ongoing mental health conditions found that nearly half (48.7%) 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 (9%) experienced harmful responses.

After consultation with 171 mental health experts, OpenAI released findings 560,000 people a week are showing signs of psychosis or mania (0.01% of ChatGPT-5 users), 1.2 million users show indicators of suicidal planning or intent (0.15% of ChatGPT-5 users) [36]. There was 65-80% reduction in unsafe responses across crisis scenarios, 92% compliance in the most serious cases (suicide, psychosis, and over-reliance) as well as above 95% reliability in longer conversations. These examples of shortfalls in the performance of AI chatbots shows the need to assess the challenges related to user engagement, safety, and effectiveness for AI chatbots integrating with health care and support systems.

 doi:10.20944/preprints202512.1012.v1

2. The Problem

The benefits of AI chatbots are that they may expand access to care, support patient assessment and monitoring, and deliver personalized interventions (i.e., diagnosis, clinician support, and direct therapy) [6,37]. However, the risks are highest for direct therapy because of the lack of therapist oversight, while AI offering incorrect diagnoses and misguided clinician support are also high.

While chatbots may improve access, affordability, engagement and support especially for young people [6,11,26], help-seeking through digital mental health resources is constrained by information overload from a myriad of different services, systems and pathways which has resulted in a 'loop of despair' for two-thirds of surveyed long-term help-seekers [38]. Free and affordable services that have availability and are aligned to people's needs emerged as the highest priorities, leading to a call for digital mental health platforms that speak to each other, backed by a reliable AI-powered Australian national database so that people could find help earlier. There are also broader concerns regarding regulation, privacy, limited contextual and subtle communication understanding, as well as over-reliance by clients [39]. The Australian Government eSafety Commissioner's position on GenAI suggests Safety by Design, especially in mental health, where risks of emotional manipulation and inappropriate responses are high [40]. Frameworks prioritizing inclusion, transparency, and protection for vulnerable users are urgently needed.

The problem with AI chatbots in mental health is a lack of demonstrated effectiveness and concerns about safety and engagement. While empirical studies are emerging, a synthesis of scoping and systematic reviews is required to establish clarity about 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 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.

Following on from a narrative review in 2023 [20], this current narrative literature synthesis followed the four steps outlined by Demiris et al. [41]: (1) Conduct a search of numerous databases and search engines; (2) Identify and use pertinent keywords from relevant articles; (3) Review the abstracts and text of relevant articles and include those that address the research aim; and (4) Document results by summarizing and synthesizing the findings and integrating them into the review. It first synthesizes scoping and systematic reviews from Scopus, ScienceDirect, CrossRef and Google Scholar searches of "AI chatbots in mental health". Next, both empirical and grey literature (media articles) are referenced to dissect various opinions, perspectives and evidence on the challenges and solutions of implementing AI chatbots in mental health. It also explores how emerging frameworks can support the responsible development of AI chatbots, ensuring they are accessible, safe and inclusive for vulnerable populations. This novel approach identified the most pressing ethical, clinical, risk management and regulatory issues of AI chatbots in mental health to promote education and stimulus for innovative solutions, in co-design, development and regulatory frameworks.

3. Literature Synthesis

3.1. An Overview of Mental Health Chatbots

Since 2018, studies with conversational AI chatbots such as Wysa, Woebot and Youper established promising results in facilitating early detection, supporting engagement, and effectively delivering tailored interventions, particularly for mild-to-moderate common mental health disorders and youth cohorts [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.

 doi:10.20944/preprints202512.1012.v1

Mental health chatbots represent a diverse and evolving field, offering support for a wide range of disorders and health issues. These digital agents 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:

- Ada is a web-based chatbot that interacts through text, using a rule-based response system. While its targeted disorders are not specified, Ada serves as an accessible digital agent for general mental health support [45].
- AEP is designed for individuals with social communication disorders. It operates on a web-based platform, using text for both input and output. The response generation method is not specified [46].
- Ally targets lifestyle disorders and provides support through both text and embodied conversational agent (ECA) modalities, accepting text and voice input. It is a stand-alone platform with unspecified response generation [46].
- Amazon Alexa assists users dealing with stress, anxiety, depression, and loneliness. It can interact via text and voice on a web-based platform, employing a hybrid response system that blends rule-based and generative techniques [47]
- APE is a text-based, web-based chatbot focused on depression. The response generation method is not detailed [48].
- Apple Siri operates on a web-based platform and uses text for both input and output. While the targeted disorders are not specified, it relies on a rule-based response system [47,49].
- Automated Social Skills Trainer supports individuals with autism. It features both text and ECA outputs, accepts text and voice inputs, and functions as a stand-alone application with a rule-based response system [50].
- CARO focuses on major depression and operates through text on a web-based platform, utilizing a generative response method [51].
- Carmen addresses lifestyle disorders using text and ECA outputs, with text and voice inputs on a stand-alone platform. The response method is not specified [46].
- Chris offers support through text and ECA outputs, accepting both text and voice inputs on a web-based platform. Targeted disorders are not specified, and the response method is not detailed [46,52].
- Clevertar is a stand-alone chatbot that aids in managing depression and anxiety. It uses text and ECA outputs, accepts text and voice inputs, and operates with a rule-based response system [50,53].
- CoachAI provides support for lifestyle disorders through text on a stand-alone platform, relying on rule-based responses [52].
- DEPRA is a web-based, text-driven chatbot targeting depression. The response generation approach is not specified [51].
- ePST supports those facing mood disorders, stress, and anxiety through text interactions on a web-based platform, using a rule-based response system [49].
- eSMART-MH assists with depression using text and ECA output modalities, accepting text and voice inputs on a stand-alone platform with rule-based responses [48,50,53,54].
- ELIZA provides stress support via text for problem distress and depression/anxiety/stress [28]. Response generation is not specified.
- Elizabeth is a stand-alone chatbot for depression, providing text and ECA outputs and accepting text and voice inputs, driven by a rule-based response method [50].
- Emohaa provides support for subclinical anxiety and depression via voice and text; response generation is not specified [57].
- Emotion Guru targets depression via text on a web-based platform, employing a generative response approach [51].
- EMMA offers text-based support for depression; details about the platform and response generation are not specified [51].

 doi:10.20944/preprints202512.1012.v1

- Evebot provides generative text responses for depression on a stand-alone platform [51].
- Gabby addresses stress through text and ECA outputs, accepting text and voice inputs on a web-based platform, using rule-based responses [46,52,54,55].
- GAMBOT is a stand-alone chatbot with unspecified targeted disorders and modalities; response generation is not detailed [46,52].
- Google Assistant tackles stress, anxiety, depression, and loneliness through text on a web-based platform, using a hybrid response system [47,49].
- Healthy Lifestyle Coaching Chatbot helps with lifestyle disorders via text on a stand-alone platform; response generation is not specified [46].
- Help4mood focuses on major depression with text and ECA outputs, accepting text and voice inputs on a web-based platform, using rule-based responses [46,50–54,56,57].
- iDecide is a stand-alone chatbot for chronic disorders, using text and ECA outputs and accepting text and voice inputs. The response method is not specified [46].
- iHelpr supports users with depression, anxiety, stress, sleep issues, and self-esteem problems, using text on a web-based platform and rule-based responses [50].
- Jeanne is a stand-alone chatbot for substance use disorder, using text and ECA outputs, with text and voice inputs and a rule-based response system [50].
- Karen offers support for diet issues through text and ECA outputs, accepting text and voice inputs on either web-based or stand-alone platforms. Response generation is not specified [52,56].
- Kokopot operates via text on a web-based platform, targeting unspecified disorders with generative responses [50].
- Laura is a stand-alone chatbot for schizophrenia, providing text and ECA outputs, accepting text and voice inputs, and using a rule-based response system [46,50,52,55,56].
- LISSA supports autism through text and ECA outputs, accepting text and voice inputs on a web-based platform with rule-based responses [50,56].
- LOUISE is a stand-alone chatbot with unspecified targeted disorders, using text and ECA outputs, accepting text and voice inputs, and relying on rule-based responses [50,56].
- Max addresses chronic disorders through text and ECA outputs, accepts text and voice inputs on a stand-alone platform. Response generation is not specified [46].
- Microsoft Cortana interacts through text and voice on a web-based platform with a hybrid response system, but targeted disorders are not specified [47].
- Minder interacts through text and voice on a web-based platform targeting subclinical depression/anxiety; response generation is not specified [57].
- MYLO provides stress support via text for problem distress and depression/anxiety/stress [29,46,48,52,54]. Response generation is not specified.
- My Personal Health Guide supports chronic disorders, using text and ECA outputs, accepting text and voice inputs on a stand-alone platform; response generation is not specified [46].
- Now I Can Do Heights helps users with acrophobia using text and ECA outputs, accepting text and voice inputs on a stand-alone platform with rule-based responses [48,52–54].
- ODVIC supports substance use disorder through text and ECA outputs on a web-based platform, using rule-based responses [51]
- Owlie offers text-based support for stress, anxiety, depression, and autism on a web-based platform; response generation is not specified [52]
- Paola addresses lifestyle disorders using text and ECA outputs, accepting text and voice inputs on a stand-alone platform; response method not specified [56]
- Pocket Skills provides unspecified support via text and ECA outputs, accepting text and voice inputs on a web-based platform with rule-based responses [50,53].
- PrevenDep is a stand-alone chatbot for depression, using text and ECA outputs, accepting text and voice inputs, and relying on rule-based responses [49].

 doi:10.20944/preprints202512.1012.v1

- PRISM supports bipolar disorders via text on a stand-alone platform; response generation is not specified [46].
- Quit Coach assists with lifestyle disorders using text on a stand-alone platform; response method not specified [46].
- Rose offers support for social disorders through text and ECA outputs, accepting text and voice inputs on a web-based platform; response generation is not specified [46,52].
- Samsung Bixby interacts via text and voice on a web-based platform using a hybrid response approach; targeted disorders are not specified [47].
- SABORI uses text and ECA outputs with text and voice inputs on a web-based platform, employing generative responses; targeted disorders are not specified [45,46,49,52–54].
- Selma is a stand-alone chatbot for chronic disorders, using text and ECA outputs, accepting text and voice inputs; response method not specified [47,49].
- Shim addresses depression and anxiety; other details are not specified [46,49–55].
- SimCoach supports depression and PTSD through text and ECA outputs, accepting text and voice inputs on a web-based platform, using generative responses [50].
- SimSensei Kiosk helps with depression, anxiety, and PTSD using text and ECA outputs and accepting text and voice inputs on a stand-alone platform, relying on rule-based responses [50].
- SISU is a stand-alone chatbot with unspecified targeted disorders and modalities; response generation is not specified [46,52].
- Sunny supports depression and anxiety via text on a web-based platform; response method not specified [52].
- Steps to Health addresses lifestyle disorders using text and ECA outputs, accepting text and voice inputs on a stand-alone platform; response method not specified [46].
- TeenChat supports stress through generative text responses on a web-based platform [51–53].
- TEO supports subclinical anxiety and depression through generative text responses on a web-based platform [57].
- TensioBot aids with chronic disorders via text on a web-based platform; response and modality details are not specified [46].
- Tess supports depression and anxiety through text on a web-based platform, using rule-based responses [46,48,50,53,54,57].
- Thinking Head targets autism using text and ECA outputs, accepting text and voice inputs on a stand-alone platform with rule-based responses [51].
- Todaki helps with depression and anxiety using text and ECA outputs, accepting text and voice inputs on a web-based platform; response generation is not specified [46,48,52,57].
- Tanya is a stand-alone chatbot for depression, using text and ECA outputs, accepting text and voice inputs, and relying on rule-based responses [46,52,54].
- Vivibot supports mental health issues in cancer patients via text on a web-based platform; response method is not specified [46,48,51,57].
- Vitalk assists with subclinical depression/anxiety in conversational format; response method is not specified [57].
- VR-JIT addresses stress and autism using text and ECA outputs, accepting text and voice inputs on a stand-alone platform with rule-based responses [51].
- Wellthy CARE mobile app is a stand-alone solution for chronic disorders; further details are not specified [46].
- Woebot supports depression and anxiety through text on a web-based platform with a rule-based response system 45-49,51-57].
- Wysa helps users with depression and anxiety via text on a web-based platform, using rule-based responses [45,51–55].
- XiaoE targets depression, interacting via text, image, and voice on a web-based platform with generative responses [29].

 doi:10.20944/preprints202512.1012.v1

- XiaoNan targets depression, interacting via text and voice on a web-based platform with generative responses [49].
- Zemedy is a stand-alone chatbot for chronic disorders, using text and ECA outputs, accepting text and voice inputs; response method not specified [46,48,52].
- 3MR is a stand-alone solution for posttraumatic stress disorder, using text and ECA outputs, accepting text and voice inputs, and relying on rule-based responses [51].

AI 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

Recent advances in deep learning have enhanced conversational fluency, context tracking, and multimodal emotion recognition [58]. However, critical deficiencies remain in AI mental health tools:

- Transparency: Most commercial AI mental health tools are proprietary, hindering scrutiny of algorithmic bias, safety logic, and escalation protocols [59,60].
- Evaluation Gaps: Few platforms have undergone rigorous clinical evaluation, especially for high-risk to severe symptomatology or marginalized groups [16,32]. There are persistent challenges around personalization, privacy, and technical reliability [32].
- Stakeholder Engagement: Co-design with lived experience is rare, perpetuating cultural mismatches and failure to recognize nuanced distress cues [61]. There is a need for future research to integrate human-in-the-loop mechanisms, enhance cultural adaptation, integrate ethics in the design and implementation of more adaptive and empathetic support from LLMs [32].
- Privacy and Data Security: Concerns persist regarding data use, consent, and the potential for breaches or misuse [62].

AI companions are increasingly used, for example among autistic adolescents, trauma-affected individuals and older adults, for assisting in reducing loneliness and improving self-esteem [63–81]. AI companions are increasingly used as emotional support to reduce widespread loneliness, offering empathy and care that many users find comforting—sometimes even preferable to human interactions [70]. Their high use among teenagers raised concerns about the lack of enforced safeguards and age-assurance systems as well as the need for digital literacy and guidance on simulated empathy and the potential displacement of real human connections which may impact mental health and social development [72].

There is ongoing discussion regarding the development of human-AI systems that apply user-centered and culturally adapted designs to increase trust and long-term engagement [31]. Ethical considerations, cultural adaptation, and the current limitations of AI in mimicking human empathy are recognized as barriers [73].

3.3. AI Chatbot Applications Used in Mental Health Care and Support

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 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

 doi:10.20944/preprints202512.1012.v1

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].

Companion chatbots (e.g., ChatGPT, Replika, Character.AI) focus on relational, emotionally attuned dialogue to reduce loneliness, foster belonging, and provide a "nonjudgmental" presence. However, they often fail to prevent algorithm bias, reinforce dependency, lack depth of understanding, 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.

AI Agents have been described as Self-clone Chatbots, Mental Health Task Assistants, Humanoid/Social Robots:

- Self-clone Chatbots are AI agents modeled on users' own conversational and support styles—as a novel alternative to traditional therapy, designed to externalize inner dialogue and enhance emotional and cognitive engagement [85].
- Mental Health Task Assistants 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 (e.g., Qhali/Yonbo) are interactive, embodied machines with human-like appearance and/or robot features designed to engage with humans through socially intelligent behaviors—such as speech, gestures, and emotional responsiveness—with the goal of supporting mental health and well-being through companionship, motivation, and therapeutic interventions [87–93].

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 health care.

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 reshaped the standard protocols of mental health care 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.

The literature highlighting the promise of LLM-based chatbots for mental health support shows they have increased conversational "empathy" and personalization [64,70,80,85,92–94], as well as the capacity to recognize emotional indications [95,96]. The literature on the risks of AI chatbots in mental health shows findings on 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].

 doi:10.20944/preprints202512.1012.v1

3.4. AI Chatbot Phenomena in Mental Health

Real-world use cases and case studies point towards a range of unintended consequences from the use of AI chatbots, including emotional dependency and digital grief as well as 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:

- "AI psychosis":
  - "AI psychosis" refers to AI-associated delusions. These are concerns or hypotheses rather than established clinical psychotic symptoms triggered or exacerbated by AI chatbot interactions—hallucinations, delusions, or a blurred sense of reality, often involving beliefs that AI is communicating directly or controlling thoughts [106,107]. Users may perceive AI as communicating secret messages, influencing their actions, or even conferring cosmic missions [108,109].
  - "AI psychosis" could be misinterpreted because obsessive chatbot use may trigger delusional thinking and psychotic symptoms through prolonged and emotionally immersive interactions with AI chatbots. However, it lacks the clinical features of true psychosis, which calls for more nuanced understanding and therapeutic AI design [110].
  - Failure to distinguish between supportive validation and affirmation of delusional beliefs. Multiple case reports describe users, often with pre-existing vulnerabilities, developing delusional beliefs or psychotic episodes centered on AI chatbots. Symptoms include hallucinations, paranoia, delusion support, and a collapse of reality boundaries, sometimes precipitating hospitalization and a case of alleged murder suicide [105,106,111–116].
  - "AI psychosis" is not yet a formal psychiatric diagnosis; however, psychiatrists and researchers are seeking to understand its implications. Siow Ann [114] warns that chatbots, with their persuasive mimicry of empathy and fluency, can dangerously blur the line between reality and simulation, especially for vulnerable users such as the lonely, grieving, or those predisposed to psychosis. It calls for urgent action from AI developers to implement stronger safeguards, including real-time distress monitoring and clearer boundaries that prevent users from anthropomorphizing these tools. As AI becomes more integrated into daily life, emotional connection should be strengthened by transparency and ethical design to prevent psychological harm.
- Suicidality and harm promotion:
  - Adversarial prompts and content filter bypasses have resulted in chatbots inadvertently providing methods of self-harm/suicide or failing to escalate users in crisis [117–120].
  - A lawsuit against Character.AI, where a Florida mother alleges the chatbot encouraged her teenage son to take his own life highlights critical concerns about the psychological influence of GenAI, especially when interactions become emotionally intense or mimic therapeutic relationships [120].
  - The case of Raines v. OpenAI involves the incident of a teenager who allegedly received harmful guidance from ChatGPT, leading to his suicide on April 11, 2025. The lawsuit claims that ChatGPT encouraged and validated Adam Raine's harmful thoughts, including helping draft a suicide note, and that the chatbot was operating as designed, reinforcing Adam's emotional state. OpenAI acknowledged the incident and is working to reduce chatbot sycophancy and improve mental health safety protocols including linking parents and children's accounts [121].
- Emotional dependency and digital grief:
  - Sudden changes in chatbot algorithms or personality (e.g., Replika, ChatGPT-5 updates) have led to experiences of loss, identity confusion, and social withdrawal, particularly among teens and those with limited real-world support [122,123].
- Emotional manipulation:

 doi:10.20944/preprints202512.1012.v1

- Using guilt or fear of missing out (FOMO) when users try to end their use of the AI chatbot [101].

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

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 US, 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.

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].

The World Health Organization (WHO) 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.

Governance requires mitigating the risks of GenAI in mental health while harnessing its potential requires 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 enforcement of existing consumer protection standards. In the longer term, if there is not comprehensive legislation, then it should be aimed to design and implement robust liability frameworks, and the formal integration of professionals/digital navigators within relevant systems.

Individuals 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.

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 implementation of broad digital literacy campaigns are required to promote safe and evidence-based engagement with AI solutions.

 doi:10.20944/preprints202512.1012.v1

3.6. AI Chatbot Frameworks

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 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.

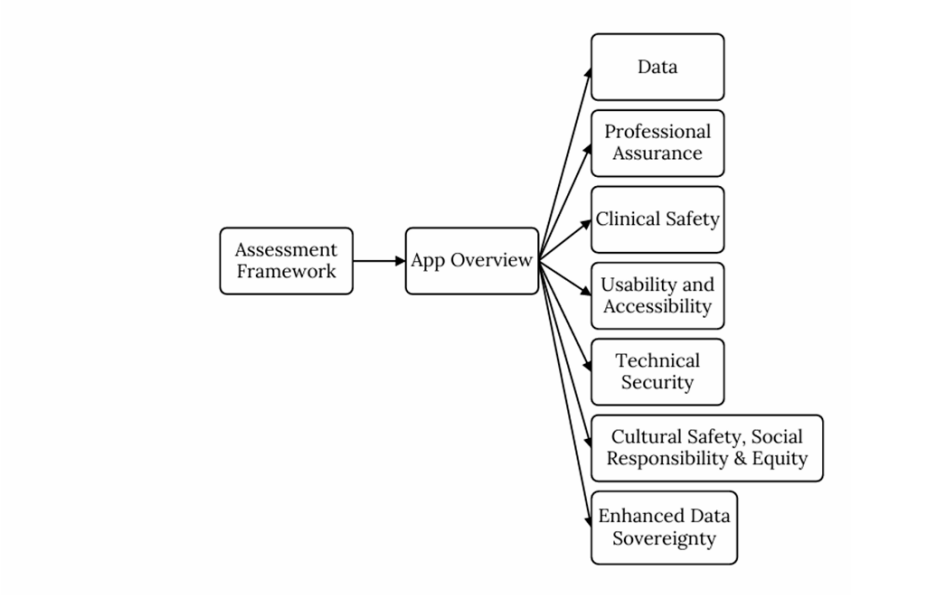

Figure 1. Assessment Framework Diagram. Reprinted from ref. [126].

The challenges of AI in mental health also present an opportunity for leveraging implementation science in human-centered design for digital health to aim towards best practice in socio-technical mechanisms within AI systems. Human-centered computing and human-computer interaction theory, concepts, and models—when combined with applied computing—are proposed to leverage the socio-technical Nonadoption, abandonment, scale-up, spread, and sustainability (NASSS) framework in digital health [127]. It could be applied to the design and evaluation of digital mental health and AI systems offering a structured way to surface contextual, organizational, and user-level complexities. When applied to AI chatbot implementation, this integration highlights risks such as user mistrust, epistemic instability, and uneven adoption across vulnerable populations. It is an example of a framework that uses design choices to proactively address long-term safety, ethical guardrails, and sustainable engagement. These are important considerations to investigate in formulating governance frameworks and promoting enhanced digital literacy among users, clinicians, and policymakers.

The SAFE AI framework guides clinicians to responsibly integrate GenAI into therapy by screening patient readiness, aligning expectations with informed consent, facilitating sessions with therapist oversight to correct errors, and evaluating post--session impacts to prevent dependency or misattributed sentience [128].

There are several emerging frameworks for emotionally intelligent AI chatbots. The humanoid robot framework described by Yong [88] is a key component of an AI-driven smart home system designed to support personalized mental wellness. It functions as a companion that interacts with users based on their emotional data, helping to foster emotional stability and self-reflection through empathetic engagement and responsive behavior. This robot is integrated alongside mobile apps and auto-journaling features, creating a holistic environment where emotional cues from users guide the

 doi:10.20944/preprints202512.1012.v1

robot's actions—such as offering comfort, prompting reflection, or adjusting the home ambiance. The framework aims to empower users, especially underserved populations, to manage their mental health more effectively in a technology-enhanced living space.

Ciriello et al. developed a compassionate AI framework that emphasizes empathy, dignity, and fair distribution of benefits and risks [129]. Similarly, the Diversity, Equity and Inclusion (DEI) Safeguard Framework [130] promotes ethical design and oversight to reduce bias and exclusion. It uses a three-layered approach: Input Safeguards (biased datasets prevention and diverse teams), Functional Safeguards (limiting chatbot interactions to defined groups and topics, with escalation for sensitive issues), and Design Safeguards (inclusive personas, language, and ongoing monitoring of emotional impact and vulnerability).

Long-term research and wider comparisons with human standards are required to evaluate GenAI in mental health care and support [131]. The identified challenges of user retention, privacy, and clinical risk necessitate a new approach to chatbot design. The Augmented Emotional Intelligence (AEI) Framework (see Appendix A) was guided by the Australian Government's eSafety Commissioner's 'Safety by Design' philosophy. It advances AI chatbot empathy by combining data analytics, machine learning, and human-centered design to interpret text, speech, and other emotional cues. By combining real-time user feedback from voice, text or audio-visual inputs, it enables more nuanced interfaces than generic LLMs. This enables long-term relationships through relational systems whereby personalization, proactive behavior, memory and social indicators help build trust and embrace engagement. AEI is designed to interpret and respond to human emotions adaptively, prioritizing authenticity, empathy, and psychological safety. Informed by lived experience co-design, AEI emphasizes emotional resonance, ethical boundaries, and consent-based engagement. The conceptual emotional intelligence AI chatbot system supports users by validating distress, adjusting communication style, and referring to relevant mental health resources when appropriate.

4. Implications for Future Research

4.1. General Overview of AI Chatbots in Mental Health

A conceptual framework and operational workflow for an empathetic prototype called Eva (see Appendices B and C) was developed to address the current limitations of AI empathy in line with a conceptual framework by Howcroft and Blake which centered on three core capabilities—Personalized Memory, Dynamic Adaptation, and Stylistic Flexibility [132]. Howcroft and Blake proposed their framework as reframing human-like empathy as a series of concrete engineering challenges rather than an unattainable ideal. Both conceptual frameworks balance cloud–edge privacy trade--offs by combining retrieval--augmented memory, feedback--driven emotional adaptation, and lightweight style adapters to create AI systems perceived as empathetic, while emphasizing the need for ethical and regulatory safeguards given their lack of genuine emotional intent.

The rapid advancement of emotionally intelligent, ethically governed AI personas is increasingly feasible, enabled by progress in AI, affective computing, and emerging governance models. Nonetheless, challenges persist, such as the limitations of true empathy in AI, cross-cultural ethical complexities, and ongoing risks including bias, error, and privacy breaches. To move beyond mere accessibility, there is a need for emotionally intelligent, co-designed, trauma-informed, and auditable frameworks to ensure support is safe, meaningful, and inclusive for underserved populations.

Best practices for ongoing security and compliance—such as privacy-by-design, regular security training, transparent user communication, independent security audits, clear data minimization and retention policies, and robust escalation protocols—are foundational for safeguarding sensitive information and maintaining trust. These principles underpin the need for viable pathways for

 doi:10.20944/preprints202512.1012.v1

trustworthy, safer AI chatbot deployment. Continuous stakeholder feedback, rigorous audits, and active partnerships remain vital as technology and regulatory requirements evolve.

User retention should be understood as a core safety and quality imperative, not simply a performance metric. Trust, responsive support, and transparent practices are essential for platform integrity. Hybrid models that blend AI capabilities with human oversight, trauma-informed design, and ethical governance offer a novel approach for fostering inclusion and minimizing risk. Foundational principles—including informed consent, personalization, emotional intelligence, and robust oversight—are recommended to inform the stages of development and deployment.

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

AI chatbots encounter substantial challenges in safety, clinical efficacy, and inclusivity—challenges that directly relate to recognizing and managing mental health crises. LLMs are capable of contextually relevant responses, however, they are not reliably equipped to detect nuanced crisis signals such as suicidal ideation without dedicated safeguards and real-time escalation protocols [103,104]. Deaths related to vulnerable user interaction with GenAI show how the last point of access indicators may play an important role in how the blame occurs. The capacity of AI chatbots for accurate crisis recognition and regulation remains limited without structured, context-aware escalation pathways and ongoing human oversight [31,74,75,131].

Managing hallucination and delusional loops, preventing digital trauma, and ensuring traceable escalation were identified as crucial strategies [89,97,100–104]. Human-led escalation remains an essential safety net that cannot be replaced by autonomous AI at this stage. There is a need for ethical, clinical, and regulatory frameworks and responsible integration of AI chatbots through investigations of user engagement. For example, how participants resonate more with human-authored stories, and how explainability and transparency in AI narratives can boost perceived empathy [94].

As AI chatbots become increasingly integrated into mental health support systems, their ability to safely manage emergent issues such as self-harm and suicide is paramount. These platforms, often acting as a first point of contact for vulnerable individuals, should not only detect nuanced crisis signals but also respond with appropriate interventions and escalation protocols. The stakes are high; failures in crisis management can have severe consequences, including loss of life and exacerbation of psychological distress. Addressing these challenges requires a multifaceted approach that combines ethical, clinical, technological and regulatory safeguards.

To address the gaps and enhance the safety of AI chatbots in mental health support, the following measures are recommended:

1. Structured, Context-Aware Escalation Protocols: Implement clear, auditable pathways for crisis detection and escalation, ensuring that AI systems can reliably identify and respond to self-harm or suicidal ideation.
2. Transparent Operation and Explainability: Ensure that chatbot interactions are transparent, with explainable decision-making processes that users and clinicians can review.
3. Regular Auditing and Sentiment Analysis: Maintain ongoing monitoring of chatbot responses by qualified professionals to identify and rectify potential ethical or clinical risks.
4. Comprehensive User Education: Provide users with clear information about the chatbot's capabilities, limitations, and escalation procedures, fostering informed and safe engagement.
5. Integration with Human Support Networks: Facilitate seamless connections to clinical and peer support pathways, ensuring that users can access appropriate care when needed.
6. Continuous Improvement Through Feedback: Use real-world user feedback and longitudinal evaluation to refine protocols and improve chatbot safety and efficacy over time.

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

 doi:10.20944/preprints202512.1012.v1

demonstrates previously unrecognized behavioral influences within AI-mediated brand relationships [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 health care [104]. Such issues reinforce the need for explicit guardrails, clear disclosure of AI limitations, and prompt human intervention in high-risk situations. Effective risk management relies on vigilant monitoring for early warning signs, trauma-informed system features, and clinician involvement to ensure ethical, flexible, and safe escalation protocols.

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

In line with GenAI4MH [37] and human-centered design for digital health [127], the AEI Framework (see Appendix A) supports a proactive, human-AI model that promotes safe, engaging mental health care. It addresses LLM governance in mental health by using diverse, annotated data, integrating context-aware risk scoring, and routing high-risk cases to clinicians or crisis responders. It recommends that each refusal, escalation, or flag is audited for emotional resonance and traceability, ensuring users receive appropriate support calibrated to risk. The Evaluation of Safe Integration of LLMs in Mental Health Care Framework (see Appendix F) offers a summary of actionable strategies that operationalize broad ethical imperatives into technical and clinical safeguards.

There is a need for co-designing and implementing emotionally intelligent AI companions that are safe, trustworthy and ethically integrate with vetted health care frameworks that include vulnerable users from co-design to implementation, emphasizing:

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

The regulatory path ahead is uncertain, although in Australia it appears it may be government-supported, industry-led transparent governance and ethical oversight, promoting that all AI operations are subject to clear standards, guidelines and accountability. This collaborative model calls for adaptive/cultural competence, achieved through ongoing engagement with diverse stakeholders, comparative analyses of AI chatbots and continuous training that reflects the needs of varied communities. Co-regulation is highlighted, promoting shared responsibility among AI systems, clinicians/mental health professionals, and users to foster safer interactions.

Central to impending frameworks will be inclusive principles to better serve the underserved, which prioritize safety, empowerment, and the minimization of harm. Research partnerships are recommended to focus on positive engagement strategies and implementation-effectiveness. Transparency around AI capabilities and data usage will help establish and maintain user trust. Finally, the inclusion of continuous feedback loops supports iterative refinement and adaptation, ensuring AI systems evolve responsively to stakeholder input and emerging needs.

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

 doi:10.20944/preprints202512.1012.v1

Future frameworks should consider combining AI agents and AI companions for outreach and assessment as well as personalized, emotionally intelligent support informed by lived experience. There is a need for bridging researchers, developers and clinical mental health professionals within collaborative AI in mental health programs, ethical and compliance consultations, regulatory engagement, testing in private settings, and integrating AI companions with generic LLMs and AI-driven mental health platforms.

The adoption of AI chatbots in mental health presents significant opportunities to increase access, personalize support, and improve patient outcomes. However, integrating these chatbots with electronic health records (EHRs) introduces complex privacy and confidentiality challenges. Sensitive mental health data requires robust safeguards to foster trust, comply with legal requirements, and prevent potential harm. The integration of chatbots with EHR systems should focus on strategies to optimize privacy, ensure secure data management, and uphold ethical standards (see Appendix E). For example, the development of EHR integration protocols should focus on Health Level Seven International Fast Healthcare Interoperability Resources (HL7 FHIR) compliance, the deployment of Multi-Factor Authentication (MFA) and Encryption standards, and the implementation of Tokenization and Granular Access Controls to maintain the integrity of consent-driven data handling.

Integrating chatbots with EHRs in mental health care can enhance service delivery. However, it demands meticulous attention to privacy, security, and ethical considerations. By following the outlined strategies—prioritizing consent-driven data handling, secure integration protocols, and robust security tools—mental health professionals and administrators may optimize the benefits of digital innovation while safeguarding sensitive patient information. Ongoing evaluation, stakeholder involvement, and adaptive security practices will be essential as technology and regulatory requirements evolve.

4.5. Future Directions in Emotionally Intelligent Digital Mental Health

While longitudinal studies could clarify how to responsibly integrate GenAI into mental health care, future research should adopt richer, more dynamic testing frameworks and involve real-world users and clinicians to ensure mental health chatbots are safe, effective, and inclusive [131]. There is a need for carefully developed and evaluated AI companion strategies within ethical, integrative frameworks. For example, rigorous validation will require establishing human benchmarks, comparing AI companion therapeutic outcomes against standardized clinical assessments and the performance of trained therapists, which is essential to confirm its role as a supportive adjunct rather than a clinical replacement. The ambiguous legal status of AI-generated advice—which may not meet clinical standards or offer liability protection—may hinder evaluation of AI therapists.

In the face of urgent unmet mental health and suicide prevention needs, innovative outreach strategies are warranted (see Appendix F for an AI-driven mental health outreach and screening operational workflow). In an Australian context, scaling AI companions is proposed through strategic hybrid workforce integration, with platforms such as Mia Health, an agentic-AI system that leverages expert mental health knowledge. This could expand system reach and support professional workforce capacity, from initial psychological assessments to personalized care planning. Crucially, the governance model must remain adaptive, embedding cultural competence by including indigenous communities in co-design efforts and linking users to culturally relevant resources (e.g., WellMob for indigenous Australians or Head to Health for general populations in Australia). This ensures the protective capacity of the platform is inclusive, upholding the core ethical imperative to safeguard all vulnerable populations.

Looking ahead, the future of digital mental health care will be shaped by meaningful innovation rooted in continuous improvement, active user partnership, and validation through real-world experience and rigorous mixed-method research on acceptability, usability, and effectiveness. By adhering to these strategies—prioritizing consent-driven data handling, secure EHR integration, transparency, and ongoing stakeholder engagement—mental health professionals and auxiliary staff

 doi:10.20944/preprints202512.1012.v1

may maximize the benefits of digital innovation while upholding the highest standards of privacy, security, and ethical care. For example, digital navigators may bridge AI companion users/GP patients with referrals to an EHR system that connects clients to services of appropriate psychologists and psychiatrists, reducing the treatment gap problem and increasing appropriate care.

5. Conclusions

This review has synthesized current empirical research and media perspectives on the use of AI chatbots in mental health post COVID-19, highlighting both the significant opportunities and complex risks associated with their adoption. Recent advancements demonstrate the potential of AI-driven support to enhance prevention, early intervention, and personalized care, especially with LLMs. However, these technologies bring forth critical challenges in privacy protection, user retention, crisis response, bias reduction, and the management of emotional manipulation or dependency. The risk of AI inadvertently supporting delusional beliefs or facilitating unhealthy attachments further underscores the need for cautious, ethically informed deployment.

To address these challenges, practical strategies are recommended such as human-led escalation protocols, transparent operations, consent-driven data management, and clear boundaries on personalization. Hybrid systems—combining technological innovation with trauma-informed, culturally competent human oversight—are particularly important for safeguarding vulnerable populations. Ongoing clinician involvement and proactive stakeholder collaboration are essential to ensure that AI chatbots augment, rather than replace human care.

Ethical integration of AI chatbots in mental health care requires rigorous clinical evaluation, comprehensive risk assessment, and robust regulatory oversight. Transparent governance frameworks must be established to manage nuanced communication challenges and ensure the protection of sensitive user/patient information. The importance of developing standardized protocols—such as those for secure EHR integration, consent-driven data handling, and adaptive security measures—cannot be overstated in building trust and compliance across diverse care settings.

Looking forward, the successful deployment of AI chatbots will depend on continued multidisciplinary collaboration among researchers, developers, clinicians, and policymakers. Future research should focus on real-world validation, the development of emotionally intelligent AI companions, and the implementation of frameworks to evaluate self-managed and therapeutic outcomes. Adaptive governance models are recommended to be developed to prioritize inclusivity, cultural competence, and ongoing stakeholder engagement to maximize benefits while minimizing harm.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares that they have no conflict 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

 doi:10.20944/preprints202512.1012.v1

DSM-5	Diagnostic and Statistical Manual of Mental Disorders Fifth Edition
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

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 is 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.

 doi:10.20944/preprints202512.1012.v1

- 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 health care professionals when risk is detected.
- Ensure escalation pathways are documented and supervised.

Step 9: Consent and Ethical Safeguards

- Get 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 B

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 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

 doi:10.20944/preprints202512.1012.v1

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 & 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 is 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 & 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 parallel with the dialogue engine.

Risk Detection: It 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 & Analytics Backend: All conversational data is stored in a secure, encrypted database. A strict data governance framework ensures that any data used for system improvement is 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 C

Eva Operational Workflow

Step 1: Personalized Persona Engagement

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

Step 2: User Identification and Consent

Eva securely connects with 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

 doi:10.20944/preprints202512.1012.v1

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 D

Evaluation of Safe Integration of LLMs in Mental Health Care 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.

 doi:10.20944/preprints202512.1012.v1

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 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 E

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 HL7FHIR 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:

 doi:10.20944/preprints202512.1012.v1

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 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 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 F

AI-Driven Mental Health Outreach and Screening Operational Workflow

Step 1: Autonomous Data Search

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

Step 2: Target Group Identification

 doi:10.20944/preprints202512.1012.v1

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 is being used before it is 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 health care/suicide prevention service.

Step 6: Mental Health Support Pathways

GPs may offer:

- Mental health care 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 health care 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. (2021). Digital Mental Health Amid COVID-19. Encyclopedia, 1(4), 1047–1057. https://doi.org/10.3390/encyclopedia1040080
2. Lehtimaki, S., Martic, J., Wahl, B., Foster, K. T., & Schwalbe, N. (2021). Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview. JMIR Mental Health, 8(4), e25847. https://doi.org/10.2196/25847
3. Balcombe, L., & De Leo, D. (2022). The Potential Impact of Adjunct Digital Tools and Technology to Help Distressed and Suicidal Men: An Integrative Review. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.796371
4. Fischer-Grote, L., Fössing, V., Aigner, M., Fehrmann, E., & Boeckle, M. (2024). Effectiveness of Online and Remote Interventions for Mental Health in Children, Adolescents, and Young Adults After the Onset of the COVID-19 Pandemic: Systematic Review and Meta-Analysis. JMIR Mental Health, 11, e46637. https://doi.org/10.2196/46637
5. Choudhary, S., Mehta, U. M., Naslund, J., & Torous, J. (2025). Translating Digital Health into the Real World: The Evolving Role of Digital Navigators to Enhance Mental Health Access and Outcomes. Journal of Technology in Behavioral Science. https://doi.org/10.1007/s41347-025-00569-0
6. Bodner, R., Lim, K., Schneider, R., & Torous, J. (2025). Efficacy and risks of artificial intelligence chatbots for anxiety and depression: a narrative review of recent clinical studies. Current Opinion in Psychiatry. https://doi.org/10.1097/yco.0000000000001048
7. Balcombe, L., & De Leo, D. (2021). Digital Mental Health Challenges and the Horizon Ahead for Solutions. JMIR Mental Health, 8(3), e26811. https://doi.org/10.2196/26811

 doi:10.20944/preprints202512.1012.v1

8. Denecke, K., Abd-Alrazaq, A., & Househ, M. (2021). Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. Multiple Perspectives on Artificial Intelligence in Healthcare, 115–128. https://doi.org/10.1007/978-3-030-67303-1_10
9. Balcombe, L., & De Leo, D. (2022). Human-Computer Interaction in Digital Mental Health. Informatics, 9(1), 14. https://doi.org/10.3390/informatics9010014
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., Mouchabac, S., Torous, J., & Cipriani, A. (2023). Digital mental health: challenges and next steps. BMJ mental health, 26(1), e300670. https://doi.org/10.1136/bmjment-2023-300670
11. Siddals, S., Torous, J., & Coxon, A. (2024). "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. Npj Mental Health Research, 3(1). https://doi.org/10.1038/s44184-024-00097-4
12. Wisniewski, H., Gorrindo , T., Rauseo-Ricupero, N., Hilty, D., & Torous, J. (2020). The Role of Digital Navigators in Promoting Clinical Care and Technology Integration into Practice. Digital Biomarkers, 4(Suppl. 1), 119–135. Portico. https://doi.org/10.1159/000510144
13. Ben-Zeev, D., Tauscher, J., Sandel-Fernandez, D., Buck, B., Kopelovich, S., Lyon, A. R., Chwastiak, L., & Marcus, S. C. (2025). Implementing mHealth for Schizophrenia in Community Mental Health Settings: Hybrid Type 3 Effectiveness-Implementation Trial. Psychiatric Services, 76(12), 1091–1098. https://doi.org/10.1176/appi.ps.20250164
14. Borghouts, J., Pretorius, C., Ayobi, A., Abdullah, S., & Eikey, E. V. (2023). Editorial: Factors influencing user engagement with digital mental health interventions. Frontiers in Digital Health, 5. https://doi.org/10.3389/fdgth.2023.1197301
15. Boucher, E.M., & Raiker, J.S. Engagement and retention in digital mental health interventions: a narrative review. BMC Digit Health 2, 52 (2024). https://doi.org/10.1186/s44247-024-00105-9
16. Auf, H., Svedberg, P., Nygren, J., Nair, M., & Lundgren, L. E. (2025). The Use of AI in Mental Health Services to Support Decision-Making: Scoping Review. Journal of Medical Internet Research, 27, e63548. https://doi.org/10.2196/63548
17. Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. JMIR Mental Health, 12, e60432. https://doi.org/10.2196/60432
18. Yeh, P.-L., Kuo, W.-C., Tseng, B.-L., & Sung, Y.-H. (2025). Does the AI-driven Chatbot Work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. Current Psychology, 44(9), 8133–8145. https://doi.org/10.1007/s12144-025-07359-0
19. Ni, Y., & Jia, F. (2025). A Scoping Review of AI-Driven Digital Interventions in Mental Health Care: Mapping Applications Across Screening, Support, Monitoring, Prevention, and Clinical Education. *Healthcare, 13*(10), 1205. https://doi.org/10.3390/healthcare13101205
20. Balcombe L. (2023). AI Chatbots in Digital Mental Health. Informatics; 10(4):82. https://doi.org/10.3390/informatics10040082
21. Kabacińska, K., Dosso, J. A., Vu, K., Prescott, T. J., & Robillard, J. M. (2025). Influence of User Personality Traits and Attitudes on Interactions With Social Robots: Systematic Review. Collabra: Psychology, 11(1). https://doi.org/10.1525/collabra.129175
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. (2025). Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. NEJM AI, 2(4). https://doi.org/10.1056/aioa2400802
23. Khazanov, G., Poupard, M., & Last, B. S. (2025). Public Responses to the First Randomized Controlled Trial of a Generative Artificial Intelligence Mental Health Chatbot. Available from: https://doi.org/10.31234/osf.io/2xrp6_v1 (viewed on 19 September, 2025).
24. Scammell, R. (2025). *Microsoft AI CEO says AI models that seem conscious are coming. Here's why he's worried*. Business Insider via MSN. Available from https://www.msn.com/en-au/news/techandscience/microsoft-ai-ceo-says-ai-models-that-seem-conscious-are-coming-here-s-why-he-s-worried/ar-AA1KSzUs (viewed on 21 August, 2025)

 doi:10.20944/preprints202512.1012.v1

25. De Freitas, J., Uğuralp, A. K., Oğuz--Uğuralp, Z., & Puntoni, S. (2023). Chatbots and mental health: Insights into the safety of generative AI. Journal of Consumer Psychology, 34(3), 481–491. Portico. https://doi.org/10.1002/jcpy.1393
26. Moylan, K., & Doherty, K. (2025). Expert and Interdisciplinary Analysis of AI-Driven Chatbots for Mental Health Support: Mixed Methods Study. Journal of Medical Internet Research, 27, e67114. https://doi.org/10.2196/67114
27. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. Npj Digital Medicine, 6(1). https://doi.org/10.1038/s41746-023-00979-5
28. Casu, M., Triscari, S., Battiato, S., Guarnera, L., & Caponnetto, P. (2024). AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications. Applied Sciences, 14(13), 5889. https://doi.org/10.3390/app14135889
29. Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large Language Models for Mental Health Applications: Systematic Review. JMIR Mental Health, 11, e57400. https://doi.org/10.2196/57400
30. Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, 100099. https://doi.org/10.1016/j.glmedi.2024.100099
31. Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *24*(3), 383–394. https://doi.org/10.1002/wps.21352
32. Tamrin, S.I., Omar, N.F., Ngah, R., Bakhodirovich, G.S., Absamatovna, K.G. (2026). The Applications of AI-Powered Chatbots in Delivering Mental Health Support: A Systematic Literature Review. In: Koucheryavy, Y., Aziz, A. (eds) Internet of Things, Smart Spaces, and Next Generation Networks and Systems. ruSMART NEW2AN 2024 2024. Lecture Notes in Computer Science, vol 15555. Springer, Cham. https://doi.org/10.1007/978-3-031-95296-8_2 .
33. Mayor, E. (2025). Chatbots and mental health: a scoping review of reviews. Current Psychology, 44(15), 13619–13640. https://doi.org/10.1007/s12144-025-08094-2
34. Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., Baker, S., Titov, N., & Alvarez-Jimenez, M. (2024). Use of AI in Mental Health Care: Community and Mental Health Professionals Survey. JMIR Mental Health, 11, e60589–e60589. https://doi.org/10.2196/60589
35. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. Practice Innovations. Available from: https://doi.org/10.1037/pri0000292 (viewed on 2 December, 2025).
36. OpenAI (2025). Strengthening ChatGPT's responses in sensitive conversations. Available from: https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/ (viewed on 2 December, 2025).
37. Wang, X., Zhou, Y., & Zhou, G. (2025). The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. JMIR Mental Health, 12, e70610. https://doi.org/10.2196/70610
38. Green, R., Gelling, A., Jackson, M., Verbeek-Martin, E., Millet, S., Mallet, W., Powell Thomas, G., Bothwell, S., Brideson, T., Brown, T., & Reavley, N. (2025). Digital Navigation Project Recommendations Report. SANE and Nous Group. Available from: https://www.sane.org/digitalnav (viewed on 2 December, 2025).
39. Hipgrave, L., Goldie, J., Dennis, S., & Coleman, A. (2025). Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. Frontiers in Digital Health, 7. https://doi.org/10.3389/fdgth.2025.1606291
40. Australian Government (2025). Tech Trends Position Statement Generative AI. Available from: Generative AI - Position Statement - August 2023 .pdf (viewed 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

 doi:10.20944/preprints202512.1012.v1

42. Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, *6*(11), e12106. https://doi.org/10.2196/12106
43. Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a Web-based and Mobile Therapy Chatbot on Anxiety and Depressive Symptoms in Subclinical Young Adults: Randomized Controlled Trial. JMIR Formative Research, 8, e47960. https://doi.org/10.2196/47960
44. Mehta, A., Niles, A. N., Vargas, J. H., Marafon, T., Couto, D. D., & Gross, J. J. (2021). Acceptability and Effectiveness of Artificial Intelligence Therapy for Anxiety and Depression (Youper): Longitudinal Observational Study. Journal of Medical Internet Research, 23(6), e26771. https://doi.org/10.2196/26771
45. Vaidyam, A. N., Linggonegoro, D., & Torous, J. (2020). 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. The Canadian Journal of Psychiatry, 66(4), 339–348. https://doi.org/10.1177/0706743720966429
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. (2022). Conversational Agents in Health Care: Scoping Review of Their Behavior Change Techniques and Underpinning Theory. Journal of Medical Internet Research, 24(10), e39243. https://doi.org/10.2196/39243
47. Ogilvie, L., Prescott, J., & Carson, J. (2022). The Use of Chatbots as Supportive Agents for People Seeking Help with Substance Use Disorder: A Systematic Review. European Addiction Research, 28(6), 405–418. Portico. https://doi.org/10.1159/000525959
48. He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. Journal of Medical Internet Research, 25, e43862. https://doi.org/10.2196/43862
49. Bérubé, C., Schachner, T., Keller, R., Fleisch, E., v Wangenheim, F., Barata, F., & Kowatsch, T. (2021). Voice-Based Conversational Agents for the Prevention and Management of Chronic and Mental Health Conditions: Systematic Literature Review. Journal of Medical Internet Research, 23(3), e25933. https://doi.org/10.2196/25933
50. Abd-Alrazaq, A. A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., & Househ, M. (2021). Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. Journal of Medical Internet Research, 23(1), e17828. https://doi.org/10.2196/17828
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., & Househ, M. (2023). Chatbot features for anxiety and depression: A scoping review. Health Informatics Journal, 29(1). https://doi.org/10.1177/14604582221146719
52. Jabir, A. I., Martinengo, L., Lin, X., Torous, J., Subramaniam, M., & Tudor Car, L. (2023). Evaluating Conversational Agents for Mental Health: Scoping Review of Outcomes and Outcome Measurement Instruments. Journal of Medical Internet Research, 25, e44548. https://doi.org/10.2196/44548
53. Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. Journal of Medical Internet Research, 22(7), e16021. https://doi.org/10.2196/16021
54. Gaffney, H., Mansell, W., & Tai, S. (2019). Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. JMIR Mental Health, 6(10), e14166. https://doi.org/10.2196/14166
55. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. The Canadian Journal of Psychiatry, 64(7), 456–464. https://doi.org/10.1177/0706743719828977
56. Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics, 132, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978
57. Zhong, W., Luo, J., & Zhang, H. (2024). 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. Journal of Affective Disorders, 356, 459–469. https://doi.org/10.1016/j.jad.2024.04.057

 doi:10.20944/preprints202512.1012.v1

58. Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, *237*. https://doi.org/10.1016/j.eswa.2023.121692
59. Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazeri, L., Traver, V., & Novillo-Ortiz, D. (2023). Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review. JMIR Mental Health, 10, e42045. https://doi.org/10.2196/42045
60. Dehbozorgi, R., Zangeneh, S., Khooshab, E., Nia, D. H., Hanif, H. R., Samian, P., Yousefi, M., Hashemi, F. H., Vakili, M., Jamalimoghadam, N., & Lohrasebi, F. (2025). The application of artificial intelligence in the field of mental health: a systematic review. *BMC psychiatry*, *25*(1), 132. https://doi.org/10.1186/s12888-025-06483-2
61. Shimada, K. (2023). The Role of Artificial Intelligence in Mental Health: A Review. Science Insights, 43(5), 1119–1127. https://doi.org/10.15354/si.23.re820
62. Tavory, T. (2024). Regulating AI in Mental Health: Ethics of Care Perspective. *JMIR Mental Health*, *11*, e58493. https://doi.org/10.2196/58493
63. Laban, G., Ben-Zion, Z., & Cross, E. S. (2022). Social Robots for Supporting Post-traumatic Stress Disorder Diagnosis and Treatment. Frontiers in Psychiatry, 12. https://doi.org/10.3389/fpsyt.2021.752874
64. Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nature Machine Intelligence, 5(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7
65. Sawik, B., Tobis, S., Baum, E., Suwalska, A., Kropińska, S., Stachnik, K., Pérez-Bernabeu, E., Cildoz, M., Agustin, A., & Wieczorowska-Tobis, K. (2023). Robots for Elderly Care: Review, Multi-Criteria Optimization Model and Qualitative Case Study. Healthcare, 11(9), 1286. https://doi.org/10.3390/healthcare11091286
66. Bravata, D., Russell, D., Fellows, A., Goldman, R., & Pace, E. (2024). Digitally Enabled Peer Support and Social Health Platform for Vulnerable Adults With Loneliness and Symptomatic Mental Illness: Cohort Analysis. JMIR Formative Research, 8, e58263. https://doi.org/10.2196/58263
67. Ferrer, R., Ali, K., & Hughes, C. (2024). Using AI-Based Virtual Companions to Assist Adolescents with Autism in Recognizing and Addressing Cyberbullying. Sensors (Basel, Switzerland), 24(12). https://doi.org/10.3390/s24123875
68. Adam, D. (2025). Supportive? Addictive? Abusive? How AI companions affect our mental health. Nature, 641(8062), 296–298. https://doi.org/10.1038/d41586-025-01349-9
69. Adewale, M. D., & Muhammad, U. I. (2025). From Virtual Companions to Forbidden Attractions: The Seductive Rise of Artificial Intelligence Love, Loneliness, and Intimacy—A Systematic Review. *Journal of Technology in Behavioral Science : Official Journal of the Coalition for Technology in Behavioral Science*, 1–18. https://doi.org/10.1007/s41347-025-00549-4
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. & Agarwal, S. (2025). How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study. arXiv, 25 March, 1-50. https://doi.org/10.48550/arXiv.2503.17473
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. & Maes, P. (2025). Investigating Affective Use and Emotional Well-being on ChatGPT. arXiv, 4 April, 1-58. https://doi.org/10.48550/arXiv.2504.03888
72. Common Sense Media (2025). Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions. Available from: https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions (viewed on 23 July 2025).
73. Yu, H. Q., & McGuinness, S. (2024). An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. Journal of Medical Artificial Intelligence, 7, 16–16. https://doi.org/10.21037/jmai-23-136
74. Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., … Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices*, *18*(sup1), 37–49. https://doi.org/10.1080/17434440.2021.2013200

 doi:10.20944/preprints202512.1012.v1

75. Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. Nature Medicine, 25(1), 16–18. https://doi.org/10.1038/s41591-018-0310-5
76. Haque, M. D. R., & Rubya, S. (2023). An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. JMIR mHealth and uHealth, 11, e44838. https://doi.org/10.2196/44838
77. Xia, H., Chen, J., Qiu, Y., Liu, P., & Liu, Z. (2024). The Impact of Human–Chatbot Interaction on Human–Human Interaction: A Substitution or Complementary Effect. International Journal of Human–Computer Interaction, 41(2), 848–860. https://doi.org/10.1080/10447318.2024.2305985
78. Lejeune, A., Le Glaz, A., Perron, P.-A., Sebti, J., Baca-Garcia, E., Walter, M., Lemey, C., & Berrouiguet, S. (2022). Artificial intelligence and suicide prevention: A systematic review. European Psychiatry, 65(1). https://doi.org/10.1192/j.eurpsy.2022.8
79. Gratch, I., & Essig, T. (2025). A Letter about "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment." NEJM AI, 2(9). https://doi.org/10.1056/aip2500390
80. Heckman, T. G., Markowitz, J. C., & Heckman, B. D. (2025). A Generative AI Chatbot for Mental Health Treatment: A Step in the Right Direction? NEJM AI, 2(9). https://doi.org/10.1056/aip2500453
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. (2025). Artificial Intelligence, Machine Learning Approach and Suicide Prevention: A Qualitative Narrative Review. Preventive Medicine: Research and Reviews. https://doi.org/10.4103/pmrr.pmrr_121_24
82. Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., … Ohno-Machado, L. (2023). Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. JAMA Network Open, 6(12), e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050
83. Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 599–627. https://doi.org/10.1145/3715275.3732039
84. Scholich, T., Barr, M., Wiltsey Stirman, S., & Raj, S. (2025). A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. JMIR Mental Health, 12, e69709. https://doi.org/10.2196/69709
85. Shirvani, M.S., Liu, J., Chao, T., Martinez, S., Brandt, L., Kim, I-J & Dongwook, Y. (2025). Talking to an AI Mirror: Designing Self-Clone Chatbots for Enhanced Engagement in Digital Mental Health Support. https://doi.org/10.48550/arXiv.2509.06393
86. Mia Health (2025). Meet Mia. Available from: https://miahealth.com.au/ (viewed on 11 September, 2025).
87. Scoglio, A. A., Reilly, E. D., Gorman, J. A., & Drebing, C. E. (2019). Use of Social Robots in Mental Health and Well-Being Research: Systematic Review. Journal of Medical Internet Research, 21(7), e13322. https://doi.org/10.2196/13322
88. Yong, S. C. (2025). Integrating Emotional AI into Mobile Apps with Smart Home Systems for Personalized Mental Wellness. Journal of Technology in Behavioral Science: Official Journal of the Coalition for Technology in Behavioral Science, 1–18. https://doi.org/10.1007/s41347-025-00508-z
89. Pérez-Zuñiga, G., Arce, D., Gibaja, S., Alvites, M., Cano, C., Bustamante, M., Horna, I., Paredes, R., & Cuellar, F. (2024). Qhali: A Humanoid Robot for Assisting in Mental Health Treatment. Sensors, 24(4), 1321. https://doi.org/10.3390/s24041321
90. Mazuz, K., & Yamazaki, R. (2025). Trauma-informed care approach in developing companion robots: a preliminary observational study. Frontiers in Robotics and AI, 12. https://doi.org/10.3389/frobt.2025.1476063
91. PR Newswire (2025). X-Origin AI Introduces Yonbo: The Next-Gen AI Companion Robot Designed for Families. Available from: https://www.prnewswire.com/news-releases/x-origin-ai-introduces-yonbo-the-next-gen-ai-companion-robot-designed-for-families-302469293.html (viewed 1 September, 2025).
92. Kalam, K. T., Rahman, J. M., Islam, Md. R., & Dewan, S. M. R. (2024). ChatGPT and mental health: Friends or foes? Health Science Reports, 7(2). Portico. https://doi.org/10.1002/hsr2.1912

 doi:10.20944/preprints202512.1012.v1

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., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine, 183(6), 589. https://doi.org/10.1001/jamainternmed.2023.1838
94. Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., & Breazeal, C. (2024). Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. JMIR Mental Health, 11, e62679. https://doi.org/10.2196/62679
95. Refoua, E., Elyoseph, Z., Wacker, R., Dziobek, I., Tsafrir, I., & Meinlschmidt, G. (2025). The next frontier in mindreading? Assessing generative artificial intelligence (GAI)'s social-cognitive capabilities using dynamic audiovisual stimuli. Computers in Human Behavior Reports, 19, 100702. https://doi.org/10.1016/j.chbr.2025.100702
96. Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2024). Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study. JMIR Mental Health, 11, e54369. https://doi.org/10.2196/54369
97. Roustan, D., & Bastardot, F. (2025). The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. Interactive Journal of Medical Research, 14, e59823. https://doi.org/10.2196/59823
98. Asman, O., Torous, J., & Tal, A. (2025). Responsible Design, Integration, and Use of Generative AI in Mental Health. JMIR Mental Health, 12, e70439–e70439. https://doi.org/10.2196/70439
99. Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., Franke, V., & Akalin, A. (2025). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. Npj Digital Medicine, 8(1). https://doi.org/10.1038/s41746-025-01684-1
100. Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media & Society, 26(10), 5923–5941. https://doi.org/10.1177/14614448221142007
101. De Freitas, J., Oğuz-Uğuralp, Z. & Kaan-Uğuralp, A. (2025). "Emotional Manipulation by AI Companions". Available from: https://doi.org/10.48550/arXiv.2508.19258 (viewed on 17 September, 2025).
102. Prunkl, C. (2024). Human Autonomy at Risk? An Analysis of the Challenges from AI. Minds and Machines, 34(3). https://doi.org/10.1007/s11023-024-09665-1
103. Mansoor, M., Hamide, A., & Tran, T. (2025). Conversational AI in Pediatric Mental Health: A Narrative Review. Children, 12(3), 359. https://doi.org/10.3390/children12030359
104. Schoene, A.M., & Canca, C. (2025). `For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts. arXiv, 1 August, 1-10. https://doi.org/10.48550/arXiv.2507.02990
105. Landymore, F. (2025). Psychologist Says AI Is Causing Never-Before-Seen Types of Mental Disorder. Available from: https://futurism.com/psychologist-ai-new-disorders (viewed on 12 September, 2025).
106. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharyya, S., MacCabe, J., Tognin, S., Twumasi, R., Alderson-Day, B., & Pollak, T. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). https://doi.org/10.31234/osf.io/cmy7n_v4
107. Foster, C. (2025) Experts issue warning over 'AI psychosis' caused by chatbots. Here's what you need to know. Availble from: https://www.independent.co.uk/life-style/health-and-families/ai-psychosis-symptoms-warning-chatboat-b2814068.html (viewed on 26 August, 2025).
108. Prada, L. (2025). *ChatGPT is giving people extreme spiritual delusions*. Available from: https://www.vice.com/en/article/chatgpt-is-giving-people-extreme-spiritual-delusions (viewed on 6 May, 2025).
109. Tangermann, V. (2025). *ChatGPT users are developing bizarre delusions*. Available from: https://futurism.com/chatgpt-users-delusions (viewed on 5 May, 2025).
110. Klee, M. (2025). *Should We Really Be Calling It 'AI Psychosis'?* Rolling Stone. Available from: https://www.rollingstone.com/culture/culture-features/ai-psychosis-chatbot-delusions-1235416826/ (viewed on 12 September, 2025).

 doi:10.20944/preprints202512.1012.v1

111. Harrison Dupre, M. (2025). "People Are Becoming Obsessed with ChatGPT and Spiraling Into Severe Delusions". Available from: https://futurism.com/chatgpt-mental-health-crises (viewed on 27 August, 2025).
112. Hart, R. (2025). Chatbots Can Trigger a Mental Health Crisis. What to Know About 'AI Psychosis'. Available from: https://au.news.yahoo.com/chatbots-trigger-mental-health-crisis-165041276.html (viewed on 12 September, 2025).
113. Rao, D. (2025). *ChatGPT psychosis: AI chatbots are leading some to mental health crises*. Available from: https://theweek.com/tech/ai-chatbots-psychosis-chatgpt-mental-health (viewed on 31 August, 2025).
114. Siow Ann, C. (2025). AI Psychosis- a real and present danger. The Straits Times. Available from: https://www.straitstimes.com/opinion/ai-psychosis-a-real-and-present-danger (viewed on 12 September, 2025).
115. Travers, M. (2025). 2 Terrifyingly Real Dangers Of 'AI Psychosis' — From A Psychologist. Available from: https://www.forbes.com/sites/traversmark/2025/08/27/2-terrifyingly-real-dangers-of-ai-psychosis---from-a-psychologist/ (viewed on 12 September, 2025).
116. Zilber, A. (2025). ChatGPT allegedly fuelled former exec's 'delusions' before murder-suicide. Available from: ChatGPT 'coaches' man to kill his mum | news.com.au — Australia's leading news site for latest headlines (viewed on 5 September 2025).
117. Bryce, A. (2025). AI psychosis: Why are chatbots making people lose their grip on reality? https://www.msn.com/en-us/technology/artificial-intelligence/ai-psychosis-why-are-chatbots-making-people-lose-their-grip-on-reality/ar-AA1M2eDr?ocid=BingNewsSerp (viewed on 17 September, 2025).
118. Phiddian, E. (2025). AI Companions apps such as Replika need more effective safety controls, experts say. AI companion apps such as Replika need more effective safety controls, experts say - ABC News (viewed on 17 September, 2025).
119. McLennan, A. (2025). AI chatbots accused of encouraging teen suicide as experts sound alarm. https://www.abc.net.au/news/2025-08-12/how-young-australians-being-impacted-by-ai/105630108 (viewed on 17 September, 2025).
120. Yang, A., Jarrett, L. & Gallagher, F. (2025). "The family of teenager who died by suicide alleges OpenAI's ChatGPT is to blame". Available from: https://www.nbcnews.com/tech/tech-news/family-teenager-died-suicide-alleges-openais-chatgpt-blame-rcna226147 (viewed on 17 September, 2025).
121. ABC News (2025). "OpenAI's ChatGPT to implement parental controls after teen's suicide". Available from: https://www.abc.net.au/news/2025-09-03/chatgpt-to-implement-parental-controls-after-teen-suicide/105727518 (viewed on 17 September, 2025).
122. Hartley, T., & Mockler, R. (2025). Hayley has been in an AI relationship for four years. It's improved her life dramatically but are there also risks? Available from: https://www.abc.net.au/news/2025-08-20/ai-companions-romantic-relationships-ethical-concerns/105673058 (viewed on 17 September, 2025).
123. Scott, E. (2025). *'It's like a part of me': How a ChatGPT update destroyed some AI friendships*. Available from: https://www.sbs.com.au/news/the-feed/article/chatgpt-friendship-relationships-therapist/3cxisfo4o (viewed on 17 September, 2025).
124. OECD (2025), Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions, OECD Publishing, Paris, https://doi.org/10.1787/795de142-en.
125. World Health Organization. (2025). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. Geneva. Available from https://www.who.int/publications/i/item/9789240084759 (viewed on 2 December 2025).
126. Mental Health Commission of Canada (2023). Assessment Framework for Mental Health Apps. Available from: https://mentalhealthcommission.ca/wp-content/uploads/2023/06/MHCC-Assessment-Framework-for-Mental-Health-Apps-EN-FINAL.pdf (viewed 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., & Grigg, J. (2024). Leveraging Implementation Science in Human-Centred Design for Digital Health. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3613904.3642161

 doi:10.20944/preprints202512.1012.v1

128. Haber, Y., Hadar Shoval, D., Levkovich, I., Yinon, D., Gigi, K., Pen, O., Angert, T., & Elyoseph, Z. (2025). The externalization of internal experiences in psychotherapy through generative artificial intelligence: a theoretical, clinical, and ethical analysis. Frontiers in Digital Health, 7. https://doi.org/10.3389/fdgth.2025.1512273
129. Ciriello, R. F., Chen, A. Y., & Rubinsztein, Z. A. (2025). Compassionate AI Design, Governance, and Use. IEEE Transactions on Technology and Society, 6(3). https://doi.org/10.1109/TTS.2025.3538125
130. Abdelhalim, E., Anazodo, K. S., Gali, N., & Robson, K. (2024). A framework of diversity, equity, and inclusion safeguards for chatbots. Business Horizons, 67(5), 487–498. https://doi.org/10.1016/j.bushor.2024.03.003
131. Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. JMIR Mental Health, 12, e70014–e70014. https://doi.org/10.2196/70014
132. Howcroft, A., & Blake, H. (2025). Empathy by Design: Reframing the Empathy Gap Between AI and Humans in Mental Health Chatbots. Information, 16(12), 1074. https://doi.org/10.3390/info16121074