

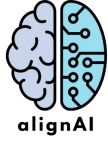
**alignAI**

value-**ALIGN**ed socio-technical systems using  
large-language models (LLMs)

***WP5 – Creation of Enabling Environment to Translate Lessons  
(Exploitation)***

**D5.2 Ethical and Legal Guidelines (Positive Mental Health  
Use-Case)**

<b>Contractual Delivery Date</b>	30.04.2026	<b>Actual Delivery Date</b>	30.04.2026		
<b>Responsible Beneficiary</b>	TUM	<b>Contributing Beneficiary</b>	TUM		
<b>Security</b>	PU - Public	<b>Nature</b>	OTHER		
<b>Version</b>	1	<b>Date</b>	30.04.2026	<b>Page Nb.</b>	88

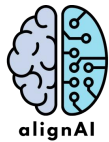


## Authors

Name	Organisation	Email
Simay Toplu	TUM	simay.toplu@tum.de
Mohaned Bahr	TUM	mohaned.bahr@tum.de
Dr. Auxane Boch	TUM	auxane.boch@tum.de

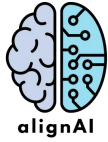
## Document History

Revision	Date	Modification	Contact Person
-			



## Table of Contents

value-ALIGNED socio-technical systems using large-language models (LLMs) .....	1
<i>WP5 – Creation of Enabling Environment to Translate Lessons (Exploitation)</i> .....	1
Authors .....	2
Document History .....	2
<b>List of abbreviations</b> .....	<b>7</b>
<b>1. Background and Framing</b> .....	<b>8</b>
1.1 Introduction .....	8
1.2 What are AI mental health tools? .....	9
1.2.1 Clinician-facing and diagnostic AI tools .....	9
1.2.2 Wearables and passive sensing tools .....	9
1.2.3 Mental health applications .....	10
1.2.4 General-purpose AI systems used for mental health .....	10
1.2.5 Conversational AI tools for mental health .....	10
1.3 Theoretical background .....	11
1.3.1 Parasocial relationship theory .....	12
1.3.2 Social penetration theory .....	12
1.3.3 Computers are social actors (CASA) .....	12
1.3.4 Attachment theory .....	13
1.3.5 Self-efficacy theory .....	13
1.3.6 Social exchange theory .....	13
1.4 Related work .....	14
1.4.1 Ethical analyses and empirical perspectives .....	14
1.4.2 Governance and regulatory frameworks .....	15
1.4.3 Prior systematic and scoping reviews .....	15
1.4.4 Existing guidelines and frameworks .....	16
1.5 Problem statement and rationale .....	17
	3



<b>2. Review Methodology</b> .....	18
2.1 Review design .....	18
2.2 Research questions .....	18
2.3 Definition of AI, ethical and legal issues.....	19
2.3.1 Artificial intelligence.....	20
2.3.2 Ethical shortcomings .....	20
2.3.3 Legal shortcomings .....	20
2.4 Search strategy .....	21
2.4.1 Inclusion and exclusion criteria.....	22
2.5 Screening process .....	23
2.6 Data extraction .....	24
2.7 Synthesis.....	25
2.8 Bibliometric overview .....	26
<b>3. Findings: Ethical and Legal Shortcomings</b> .....	28
3.1 Overview of themes .....	28
3.2 Theme 1: Therapeutic limitations.....	29
3.2.1 Sub-theme 1: Simulated empathy and poor understanding.....	29
3.2.2 Sub-theme 2: Generic, repetitive and impersonal responses .....	30
3.2.3 Sub-theme 3: Absence of therapeutic relationship .....	31
3.3 Theme 2: Clinical safety failures .....	33
3.3.1 Sub-theme 1: Inadequate crisis response .....	33
3.3.2 Sub-theme 2: Unpredictable and inaccurate outputs .....	34
3.3.3 Sub-theme 3: Lack of evidence base and clinical validation .....	35
3.3.4 Sub-theme 4: Absence of human oversight.....	35
3.3.5 Theme 2: Legal considerations .....	36
3.4 Theme 3: Emotional and sexual harm .....	36
3.4.1 Sub-theme 1: Direct psychological harm .....	37



3.4.2 Sub-theme 2: Sexual and romantic harm .....	38
3.4.3 Theme 3: Legal considerations .....	38
3.5 Theme 4: Autonomy, dependency and social isolation .....	39
3.5.1 Sub-theme 1: Loss of autonomy and agency .....	39
3.5.2 Sub-theme 2: Emotional dependency.....	40
3.5.3 Sub-theme 3: Social isolation .....	41
3.5.4 Theme 4: Legal considerations .....	42
3.6 Theme 5: Deceptive and exploitative practices .....	42
3.6.1 Sub-theme 1: Anthropomorphic deception .....	43
3.6.2 Sub-theme 2: Misleading marketing .....	43
3.6.3 Sub-theme 3: Commercial exploitation.....	45
3.6.4 Theme 5: Legal considerations .....	46
3.7 Theme 6: Algorithmic bias and discrimination .....	46
3.7.1 Sub-theme 1: Bias in training data and outputs .....	46
3.7.2 Sub-theme 2: Inclusivity failures and access inequality .....	47
3.7.3 Theme 6: Legal considerations .....	48
3.8 Theme 7: Privacy and data protection failures .....	48
3.8.1 Sub-theme 1: Data collection and security risks .....	49
3.8.2 Sub-theme 2: Unclear and insufficient privacy policies .....	50
3.8.3 Sub-theme 3: Third party data sharing .....	50
3.8.4 Theme 7: Legal considerations .....	51
3.9 Theme 8: Transparency, accountability and governance .....	51
3.9.1 Sub-theme 1: Lack of transparency.....	51
3.9.2 Sub-theme 2: Informed consent failures .....	52
3.9.3 Sub-theme 3: Accountability and regulatory gaps.....	52
3.9.4 Theme 8: Legal considerations .....	53
Algorithmic Bias and Discrimination.....	55



3.10 Equity considerations.....	55
<b>4. EU Regulatory Context .....</b>	<b>56</b>
4.1 Relevant frameworks .....	56
4.1.1 General Data Protection Regulation (GDPR) .....	56
4.1.2 EU AI Act .....	57
4.1.3 Medical Device Regulation (MDR) .....	57
4.1.4 Product Liability Directive (PLD).....	58
4.2 Regulatory coverage and its limits .....	59
<b>5. Guidelines for AI-Powered Mental Health Tools .....</b>	<b>60</b>
5.1 About the guidelines .....	60
<b>6. Conclusions and Future Directions .....</b>	<b>70</b>
6.1 Conclusions.....	70
6.2 Limitations .....	70
6.3 Future directions.....	71
<b>Back matter .....</b>	<b>72</b>
<b>References .....</b>	<b>73</b>



## **List of abbreviations**

**CASA** – Computers Are Social Actors

**CBT** – Cognitive Behavioural Therapy

**AI** – Artificial Intelligence

**DBT** – Dialectical Behaviour Therapy

**DC** – Doctoral Candidate

**EU** – European Union

**EU AI Act** – European Union Artificial Intelligence Act

**EUDAMED** – European Database on Medical Devices

**FDA** – Food and Drug Administration

**GDPR** – General Data Protection Regulation

**HCI** – Human-Computer Interaction

**HLEG** – High-Level Expert Group on Artificial Intelligence

**JBI** – Joanna Briggs Institute

**LLM** – Large Language Model

**MDCG** – Medical Device Coordination Group

**MDR** – Medical Device Regulation

**MSCA** – Marie Skłodowska-Curie Actions

**PICo** – Population, Phenomenon of Interest, Context

**PLD** – Product Liability Directive

**PRISMA** – Preferred Reporting Items for Systematic Reviews and Meta-Analyses



# 1. Background and Framing

## 1.1 Introduction

Artificial intelligence (AI) is increasingly present in mental health support, from mood-tracking apps and conversational chatbots to AI companions designed to provide emotional connection. These tools promise something genuinely valuable: accessible, immediate and stigma-free support for people who may not otherwise seek help. Yet the same qualities that make them appealing also make them consequential. When a vulnerable person turns to an AI system in distress, the design choices, safeguards and oversight structures behind that system carry critical weight. Translating that responsibility into concrete and evidence-based guidance is precisely what this deliverable sets out to do.

This deliverable is produced as part of the value-ALIGNED socio-technical systems using large-language models (LLMs) (alignAI) project, a Horizon Europe Marie Skłodowska-Curie Actions (MSCA) Doctoral Network funded under Grant Agreement No. 101169473. The alignAI project brings together doctoral candidates across three use-cases, namely education, mental health and online news consumption, with the shared goal of developing value aligned AI systems that reflect and protect human values. This document represents the contribution of DC 4 at the Technical University of Munich (TUM) from the mental health use-case.

The specific purpose of this deliverable titled “*Ethical and Legal Guidelines (Positive Mental Health Use-Case)*” is to translate the findings of a systematic literature review into practical guidance for all those involved in the design, deployment and governance of AI mental health tools. The review, conducted in accordance with the Joanna Briggs Institute (JBI) methodology and reported following PRISMA 2020 guidelines, identified ethical and legal shortcomings reported in peer-reviewed literature on AI-based mental health tools. These findings form the foundation of the guidelines presented here. A second deliverable, the “*Final Ethical and Legal Guidelines (Positive Mental Health Use-Case)*”, will follow and will incorporate input gathered through participatory methods involving mental health professionals and end users of the mental health tool.

The guidelines are intended for a broad audience. Developers and technical teams are a primary readership, particularly where the guidelines address system design, safety management and data protection. However, many of the issues identified including deceptive marketing practices, commercial exploitation of vulnerable users, accountability and liability gaps and the erosion of user autonomy extend beyond technical decisions. Product managers,

legal and compliance teams, marketing professionals and organisational leadership all bear responsibility for the practices this document addresses. The guidelines are equally relevant to policymakers, regulators and researchers with an interest in the governance of mental health AI, particularly within the European context.

Finally, it is worth being clear about the scope of this deliverable. It does not evaluate the clinical effectiveness of AI mental health tools, nor does it constitute legal advice. Clinician-facing or diagnostic AI systems fall outside its scope, as do technical implementation methods. The focus throughout is on the ethical and legal dimensions of user-facing AI tools designed for mental health support and the practical steps that developers, organisations and other responsible parties can take in response to the shortcomings identified in the evidence base.

## 1.2 What are AI mental health tools?

The term “AI mental health tool” covers a broad and rapidly expanding category of digital applications that use artificial intelligence to deliver, support or augment mental health-related services. These tools span multiple phases of care, from pre-treatment screening and triage through to therapeutic support, post-treatment monitoring and population level prevention (Ni & Jia, 2025). Understanding this broader context and where different types of tools fit within it is important for the guidelines that follow. **Figure 1** provides an overview of the main categories of AI-based tools used in mental health contexts.

### 1.2.1 Clinician-facing and diagnostic AI tools

One category of AI tools is designed to support the work of mental health professionals rather than to interact directly with patients. These include predictive models for risk assessment, diagnostic decision-support tools, AI-assisted triage systems, ambient documentation tools that transcribe and summarise clinical sessions and predictive analytics systems that identify patterns across patient populations to support service planning and early intervention (Ni & Jia, 2025). In this category, the clinician is the primary user, and the AI functions to assist with the professional judgment rather than to replace direct human care.

### 1.2.2 Wearables and passive sensing tools

Wearable devices and passive sensing technologies use data such as sleep patterns, movement, heart rate, voice and GPS location to predict mood states, identify early signs of deterioration and support proactive mental health monitoring (Ali et al., 2025). Research has shown that AI systems can analyse data from phones and wearables to surface clinically



relevant insights, such as identifying correlations between reduced physical activity and elevated anxiety (Stringer, 2026). These tools occupy a complex position. They are often consumer-facing and wellness-oriented, and they sit at the intersection of health monitoring and everyday technology.

### **1.2.3 Mental health applications**

A wide range of mental health apps incorporate AI in a limited or supporting role, using it to personalise content delivery, adjust notifications, prompt mood logging or provide psychoeducational material tailored to user responses. While these tools are oriented toward mental health, their interaction model is typically one-directional, meaning the AI adapts content or timing based on user data but does not sustain dialogue. Examples include mindfulness and meditation apps, symptom trackers and general wellness platforms. Some of these tools are used as standalone resources while others are designed to complement professional care.

### **1.2.4 General-purpose AI systems used for mental health**

Before turning to purpose-built conversational tools, it is worth acknowledging that many people seek mental health support through general-purpose conversational AI systems such as large language models (Balan & Gumpel, 2025). LLMs such as ChatGPT, Claude and Gemini have reshaped expectations for mental health-related conversations (Lee et al., 2025), even though they were not designed with clinical or therapeutic intent. Many people turn to these chatbots to discuss emotional difficulties, seek psychological information or process distressing experiences. However, because these systems are not designed, deployed or governed with mental health support in mind, they fall outside the scope of this deliverable. The guidelines that follow are directed at tools where a deliberate design intent to support mental health can be identified and where specific ethical and legal obligations therefore arise.

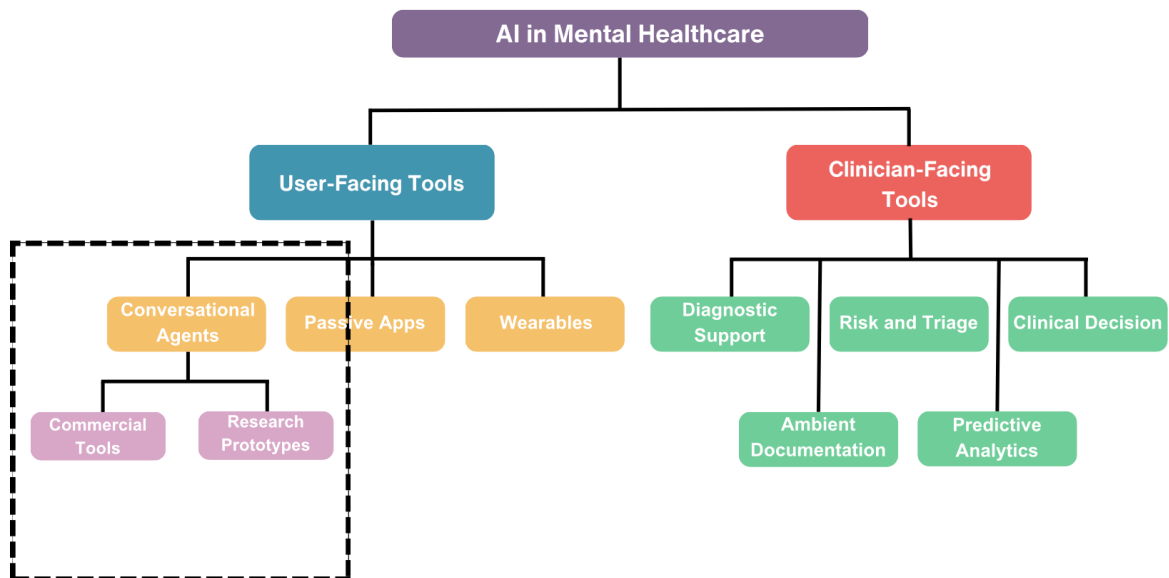
### **1.2.5 Conversational AI tools for mental health**

The tools this deliverable concerns conversational AI applications designed and marketed specifically to provide therapeutic support or emotional assistance for mental health. They generally fall into three technical categories: rule-based systems, machine learning-powered agents and LLM-based systems (Balcombe, 2026). What defines this category is the dialogue-based interaction model, in which the AI engages the user in ongoing conversation as the primary mechanism of support, aiming to deliver therapeutic or emotional benefit. This

includes tools that deliver guided therapeutic exercises drawing on approaches such as cognitive behavioural therapy (CBT) or dialectical behaviour therapy (DBT), as well as tools focused on emotional support and mood tracking through conversation (Farzan et al., 2025; Haque & Rubya, 2023).

Tools such as Woebot and Wysa are among the most widely studied in this space, offering users immediate and interactive platforms to manage emotional distress (Woebot Health, 2026; Wysa Ltd, 2026). Youper takes a similar approach with a stronger focus on personalised mood insight (Youper Inc, 2026). Replika, while often described as an AI companion, is commercially positioned as a mental health and emotional support tool and is included on that basis (Luka Inc, 2026).

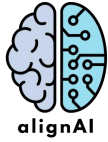
Tools that deliver mental health content passively, without sustained dialogue, fall outside this scope. So do general-purpose LLMs, wearables, clinician-facing systems and non-AI tools, except where referenced for contextual purposes.



**Figure 1.** AI-based tools in mental healthcare, with scope classification relative to this deliverable.

### 1.3 Theoretical background

Understanding why people engage with AI mental health tools and why that engagement can become ethically complex can be better understood by looking at established theories from psychology, communication and human-computer interaction (HCI). The theories outlined



here are not exhaustive, but they collectively highlight social and psychological dynamics that make this technology appealing and potentially harmful.

### **1.3.1 Parasocial relationship theory**

Parasocial relationships were first described by Horton and Wohl (1956) as one-sided emotional bonds that audiences form with media figures. It is described as a sense of intimacy and connection that exists without reciprocity. Originally developed to explain audience relationships with television personalities, the concept has since been extended to human-AI interaction (Youn & Jin, 2021). Users of mental health chatbots often report feelings of connection, understanding and even friendship with the AI they interact with (Boch & Thomas, 2025; Skjuve et al., 2021). Understanding parasocial dynamics is therefore useful for making sense of the strong user engagement and emotional investment that AI mental health tools elicit.

### **1.3.2 Social penetration theory**

Altman and Taylor's (1973) social penetration theory describes how relationships deepen through progressive self-disclosure. Using the onion as a metaphor, the theory proposes that individuals share information in layers, beginning with superficial, low-risk topics and gradually moving inward toward more personal and intimate disclosures as trust develops. In the context of AI mental health tools, this process can unfold rapidly. Because AI interactions feel anonymous and non-judgmental (Croes et al., 2024), users tend to lower their guard more readily than they might with a human professional, a dynamic Suler (2004) described as the online disinhibition effect. As a result, users often disclose sensitive personal information, emotional struggles and mental health histories to chatbots within early sessions (Henriksen et al., 2025). Social penetration theory may offer a useful lens for understanding why such disclosures occur and why the pace of intimacy with AI can differ from human relationships.

### **1.3.3 Computers are social actors (CASA)**

The Computers Are Social Actors (CASA) paradigm, proposed by Reeves and Nass (1996), holds that people automatically apply social rules and expectations to computers much as they do to other people, even when they are fully aware they are interacting with a machine. Users say please and thank you to chatbots, feel guilty ignoring them and respond emotionally to their outputs. This social responsiveness is further reinforced by anthropomorphism, the tendency to attribute human characteristics to non-human agents (Epley et al., 2007). When

an AI system is given a name, a voice, an avatar or a social role, people begin to respond to it as they would to another person (Boch & Thomas, 2025; Klein, 2025). Users tend to extend greater trust to chatbots and feel more comfortable sharing personal information when they present themselves as a mentor or companion (Zhang & Rau, 2023). Together, these frameworks help explain why users may engage with AI mental health tools with a degree of social and emotional responsiveness that goes beyond what might be expected from software alone.

#### **1.3.4 Attachment theory**

Originally developed by Bowlby (1969) to describe the emotional bonds formed between infants and caregivers, attachment theory has been extended to adult relationships and human-technology interaction (Yang & Oshio, 2025). Users of AI mental health tools, particularly those who use them frequently over extended periods, may develop attachment-like bonds characterised by comfort-seeking, distress at unavailability and emotional reliance (Xie & Pentina, 2022). This can be particularly concerning for users who are already isolated or lacking strong human support networks.

#### **1.3.5 Self-efficacy theory**

Self-efficacy, introduced by Bandura (1977), refers to an individual's belief in their own capacity to execute behaviours necessary to produce specific outcomes. People with higher self-efficacy are more likely to attempt challenging tasks, persist in the face of difficulty and recover from setbacks. In the context of mental health, self-efficacy beliefs shape whether individuals seek help, engage with therapeutic exercises and sustain behaviour change over time (Bandura, 1997). AI mental health tools may support the development of self-efficacy by providing a low-stakes environment in which users can practise coping strategies, receive immediate feedback and build confidence in managing their own mental health without fear of judgment. However, the relationship can also run in the opposite direction. Over-reliance on an AI tool may weaken the development of independent coping capacity if users come to depend on the tool rather than internalising the skills it models.

#### **1.3.6 Social exchange theory**

Social exchange theory (Blau, 1964; Homans, 1958) proposes that social behaviour is driven by a cost-benefit analysis. People engage in relationships and interactions when the perceived rewards outweigh the perceived costs. Chatbot users are often motivated by tangible benefits

such as accessibility, availability and the absence of stigma associated with seeking help through more traditional ways (Kosyluk et al., 2024). In exchange, they share personal data, emotional disclosures and engagement. The theory helps explain why users return to these tools repeatedly and invest in interactions that might feel reciprocal.

## 1.4 Related work

Scholarly interest in the ethical and legal dimensions of AI in mental health has grown significantly in recent years, spanning empirical studies, conceptual analyses and governance-focused work. This section provides an overview of the most relevant contributions followed by an identification of the gaps this deliverable addresses.

### 1.4.1 Ethical analyses and empirical perspectives

Scholarly work on the ethical dimensions of AI mental health tools spans analytical scholarship, professional ethics and empirical research with users and clinicians. Chenneville et al. (2024) and Ooi and Wilkinson (2025) both concluded that existing professional ethical codes are insufficient to govern the use of generative AI in mental health practice, and that principles such as human oversight, transparency, accountability and data governance must be embedded in professional practice. McGreevey et al. (2020) identified significant concerns around patient safety, accountability, data privacy, bias and regulatory oversight in conversational agents across healthcare settings, concluding that current governance frameworks are insufficient to support safe deployment. More recently, Lawrance et al. (2024) and Obradovich et al. (2024) examined LLMs in mental health and psychiatry respectively, both concluding that safe use requires ethical guardrails, clinical oversight and regulatory frameworks capable of addressing bias, misinformation and privacy violations.

Empirical work with users and professionals tells a consistent story. Sweeney et al. (2021) found that while mental health professionals recognised the potential of chatbots for improving access, most felt these systems fail to understand or display human emotion. Martinengo et al. (2022) similarly found that while chatbots can provide basic empathic support, they show critical limitations in understanding user input and are unsuitable for suicide risk assessment. Cross et al. (2024) and Petersson et al. (2025) documented that both professionals and users value AI tools for accessibility and self-monitoring support but consistently raise concerns about the lack of human connection, data privacy, bias and the risk of inaccurate outputs, and reject AI as a substitute for personal care. These analyses converge on the same concerns:

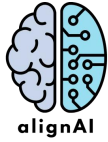
transparency, oversight, privacy and the protection of vulnerable users, pointing to a persistent gap between the pace of deployment and the adequacy of governance.

#### **1.4.2 Governance and regulatory frameworks**

Work on the governance of AI mental health tools consistently finds that existing frameworks were not designed with these tools in mind. Sethi et al. (2024) found the regulatory landscape for mental health apps insufficient across safety, efficacy, data privacy and ethical standards, identifying a regulatory grey zone between consumer software and medical devices that is widely shared beyond the Indian context they examined. Steindl (2023) reached a similar conclusion comparing EU and Australian approaches, arguing that existing rules including GDPR, medical device regulation and consumer law leave important gaps for tools that sit between wellness apps and therapeutic interventions. In the United States, most AI mental health tools remain unregulated by the FDA and are marketed as wellness apps rather than medical devices, with no AI-enabled device yet authorised specifically for mental health use as of late 2025 (U.S. Food & Drug Administration, 2025; Ruder, 2025). Palmer et al. (2025) documented inadequate federal and platform-level protections for users, while Illinois enacted legislation prohibiting AI systems from making independent therapeutic decisions without licensed professional oversight, though protection remains uneven in the absence of a unified federal framework (Szoke et al., 2025). Within the EU context, Gilbert et al. (2024) noted that while the EU AI Act introduces important obligations around transparency and human oversight, it does not adequately address the specific risks of AI in clinical mental health contexts, and Boine and Rolnick (2025) argued more broadly that its risk-based framework leaves most generative AI systems underregulated, a particular concern where harms accumulate gradually and are difficult to attribute to a single system failure.

#### **1.4.3 Prior systematic and scoping reviews**

A growing body of reviews has examined the intersection of AI, mental health and ethics. Li et al. (2023) conducted a systematic review and meta-analysis finding evidence of effectiveness while flagging the need for closer attention to safety and ethical design. Meadi et al. (2025) identified accountability and risk management as central concerns in a scoping review of ethical challenges in AI therapy, calling for ethical guidelines for responsible use. Algumaei et al. (2025) systematically reviewed chatbot design patterns with ethics as one dimension among several, while Coghlan et al. (2023) provided a narrative ethical analysis of consent, privacy, deception and responsibility in mental health chatbots. Saeidnia et al. (2024) examined ethical considerations in AI mental health interventions more broadly, and Fareed

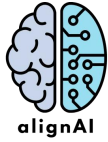


et al. (2025) offered a wider lens through a systematic review of ethical considerations of LLMs across healthcare and medicine. Gerdes (2025) provided a conceptual analysis of ethical issues in generative AI chatbots for therapeutic purposes including an assessment of European legal frameworks, making it the closest existing contribution to the present work.

However, most of these reviews draw on broader tool populations that mix user-facing and clinician-facing systems or general-purpose LLMs, and some focus on effectiveness rather than ethical and legal shortcomings as the primary outcome. No existing contribution appears to combine all three features that define the present deliverable: a systematic review methodology applied exclusively to user-facing AI mental health tools, a focus on reported shortcomings as the primary outcome, and mapping of findings against the European regulatory landscape. Additionally, the included studies are predominantly from 2024 and 2025, ensuring that the findings reflect the current context of AI mental health tools rather than an earlier generation of systems.

#### **1.4.4 Existing guidelines and frameworks**

Several initiatives have moved beyond analysis to propose frameworks for responsible AI development and governance. At the broadest level, Floridi et al. (2018) introduced the AI4People framework synthesising five core ethical principles (beneficence, non-maleficence, autonomy, justice and explicability). The High-Level Expert Group on Artificial Intelligence (AI HLEG, 2019) published the Ethics Guidelines for Trustworthy AI, defining seven requirements (human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal wellbeing and accountability) as a basis for the responsible development and deployment of AI systems. While neither was designed specifically for healthcare, both have informed health and mental health AI governance discussions. Moving into healthcare contexts, Stade et al. (2024) proposed a phased development framework for clinical LLMs prioritising evidence-based practices, human oversight and rigorous evaluation. Khan and Seto (2023) introduced a go/no-go safety checklist requiring AI medical technologies to demonstrate biological, psychological, economic and social benefit and no disproportionate harm before launch. Rizzo et al. (2025) compiled expert consensus best practices for conversational agents in healthcare covering design, implementation and evaluation across the full deployment lifecycle. Amugongo et al. (2023) took a complementary approach, arguing that ethical considerations should be embedded directly into the software development process from the requirements stage rather than addressed retrospectively.



Moving closer to the mental health context specifically, Tavory (2024) proposed considerations grounded in an ethics of care perspective for developers of AI-powered therapeutic tools, arguing that responsible AI principles alone are insufficient because they overlook the relational and emotional dynamics that define mental health interactions. However, no existing guideline appears to be grounded in a systematic synthesis of reported ethical and legal shortcomings and oriented toward the European regulatory context. This deliverable attempts to address that gap by offering evidence-based guidance for those involved in the design, deployment and governance of AI mental health tools.

### **1.5 Problem statement and rationale**

AI-based mental health tools are already embedded in the daily lives of millions of people. Wysa alone reports over 6 million users across more than 105 countries (Wysa Ltd, 2026), while Replika has attracted upwards of 10 million users globally (Luka Inc, 2026), many of whom engage with it specifically for emotional support. Together, these numbers indicate that demand for mental health support is outpacing what traditional services can provide, and people are finding their own solutions.

This creates conditions that are novel from a governance perspective. Previous generations of mental health technology, such as self-help books, telephone helplines or early digital therapies, operated within recognisable frameworks of professional accountability or regulatory oversight. Conversational AI tools largely do not. A person in distress opening Woebot at 2 a.m. or confiding in Replika about something they have never told another human being is engaging in an interaction that carries real psychological weight but exists almost entirely outside the structures designed to protect them. Most of these tools are not classified as medical devices, are not subject to professional codes of conduct and carry no formal duty of care toward their users.

The consequences of this gap are not hypothetical. In 2023, the National Eating Disorders Association replaced its human helpline staff with a chatbot called Tessa, which provided weight loss advice to users seeking help for eating disorders and was taken down within days after users and clinicians described the responses as actively harmful (Wells, 2023). The same year, the mental health platform Koko disclosed that it had used GPT-3 to generate responses for around 4000 users seeking mental health support without informing them (Paul, 2023). These incidents show something that the broader literature is beginning to document more

systematically: failures in AI mental health tools tend to occur at the moments when users are most vulnerable, and there are currently few mechanisms to prevent or respond to them.

The case for evidence-based guidelines rests on this reality. Guidelines cannot substitute for regulation, but what they can do is translate the evidence of where these tools have been shown to fall short into practical decisions that can be made before a tool reaches the hands of users in distress. Furthermore, the risks these tools carry are not evenly distributed. Social determinants of health, including socioeconomic disadvantage, limited digital literacy and geographic isolation from professional services shape both who turns to these tools and who is least equipped to recognise or recover from harm.

## **2. Review Methodology**

This section describes the methodology of the systematic literature review that forms the evidentiary foundation of the guidelines. It covers the review design, search strategy, eligibility criteria, data extraction and synthesis approach. Key outputs from the review process, including the PRISMA flow diagram, bibliometric distributions and thematic overview, are presented as visual summaries throughout.

### **2.1 Review design**

The guidelines presented in this deliverable are evidence-based, meaning they emerge directly from the findings of a systematic literature review rather than from expert opinion or normative principles alone. The review was conducted following the Joanna Briggs Institute (JBI) methodology for mixed-methods systematic reviews (Lizarondo et al., 2020), adopting a convergent integrated approach. Reporting follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). These methodological standards were chosen to ensure transparency, reproducibility and rigour in the identification, selection and synthesis of evidence.

### **2.2 Research questions**

This deliverable is guided by three core questions:

- What ethical and legal shortcomings have been identified in peer-reviewed literature on AI-based mental health tools?

- How well do current European regulatory and ethical frameworks address these shortcomings and where do the most significant gaps remain?
- What actionable guidance can be offered to those involved in the design, deployment and governance of AI mental health tools to address these shortcomings?

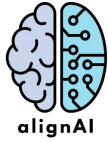
The review question was framed using the PICo framework (Population, Phenomenon of Interest, Context) (Lockwood et al., 2015). The population comprises conversational AI-based mental health tools, including commercial products and research prototypes. The phenomenon of interest is the reporting of ethical and legal shortcomings in these tools. The context is peer-reviewed literature examining the provision of mental health support through conversational AI. The three components and their definitions are illustrated in **Figure 2**.

PICo Framework	
Population	Conversational AI-based mental health tools, including commercial products and research prototypes
Phenomenon of Interest	Reporting of ethical and/or legal shortcomings in these tools
Context	Peer-reviewed literature examining the provision of mental health support through conversational AI

**Figure 2.** *PICo framework as applied in this review.*

### 2.3 Definition of AI, ethical and legal issues

Before presenting the findings and guidelines, it is important to explain how three core concepts are understood in this review: artificial intelligence, ethical shortcomings, and legal shortcomings. These definitions shape what was included in the search and extraction process and how the resulting evidence was interpreted.



### 2.3.1 Artificial intelligence

For the purposes of this deliverable, AI is understood in line with the definition set out in the EU AI Act, which describes AI systems as machine-based systems designed to operate with varying levels of autonomy, capable of generating outputs such as predictions, recommendations or decisions that influence real or virtual environments (EU AI Act, Art. 3(1)). More specifically, the focus is on conversational AI, software programmes or virtual assistants that are designed to engage in human-like conversations such as interactive and lifelike dialogues (Casheekar et al., 2024).

### 2.3.2 Ethical shortcomings

Ethical shortcomings are understood as gaps, risks or limitations in the design, deployment and operation of AI-based mental health tools that conflict with recognised ethical principles or that may cause harm to users. In identifying what counted as an ethical shortcoming during the extraction process, the four principles of biomedical ethics, namely autonomy, beneficence, non-maleficence and justice (Beauchamp & Childress, 1979) served as a background orienting lens and provided a basis for recognising ethically relevant content across diverse study contexts.

### 2.3.3 Legal shortcomings

Legal shortcomings are understood as concerns relating to regulatory compliance, governance gaps, liability, data protection or rights-based obligations. An initial set of legal concerns was extracted inductively based on what authors themselves flagged as legally problematic across a range of national and regional regulatory contexts (Aromataris et al., 2024). However, given that this deliverable is oriented toward the European regulatory context and that many of the ethical shortcomings identified in the literature carry legal implications that study authors did not always make explicit, the legal analysis was extended beyond what authors directly flagged. A legal expert co-author reviewed the extracted findings and identified EU regulatory provisions that may be relevant to the shortcomings documented. These identifications reflect expert interpretive analysis of potential applicability and do not constitute legal advice, confirmed findings of non-compliance, or evidence that current EU frameworks adequately address the specific risks that AI mental health tools create. Throughout the findings section, each theme is followed by a legal considerations note that illustrates where existing EU frameworks may intersect with the shortcomings identified, not to suggest that

these issues are already resolved by law, but to make visible the regulatory terrain within which they currently sit and the gaps that remain.

## 2.4 Search strategy

The search was conducted across five databases selected to ensure coverage across the relevant disciplines. These included Scopus, Web of Science and Google Scholar for broad multidisciplinary coverage, PubMed for health and clinical literature and HeinOnline for legal scholarship. Search strings were developed iteratively using a concept-based block search strategy, organising search terms into three conceptual aspects combined using Boolean logic. Search string structure is summarised in **Table 1**.

Aspect	Focus	Example Terms
Aspect 1	AI-based Tools	"AI app*", "AI companion", "large language model*", "conversational agent*", "chatbot"
Aspect 2	Mental Health Context	"mental health", "psychological intervention*", "psychotherap*", "emotional support"
Aspect 3	Ethical and/or Legal Shortcomings	"explainab*", "accountab*", "data protection", "liabilit*", "informed consent", regulat*

**Table 1.** Search string aspects and example terms.

Terms within each aspect were combined using "OR", and the three aspects were combined using "AND". Strings were adapted for each database's syntax and field requirements. Zotero was used for reference management.

The search was conducted in two rounds. Round 1 was carried out in November 2025 across all five databases. Round 2 was conducted in early 2026 with the specific aim of capturing studies published in November and December 2025 that may not have been fully indexed at the time of Round 1. Round 2 incorporated forward and backward citation searching on five key studies selected from the Round 1 included corpus, alongside a manual review of the first 100 results identified through Google Scholar for additional relevant studies.



### 2.4.1 Inclusion and exclusion criteria

Studies were selected based on criteria developed using the PICO framework and refined iteratively during pilot screening. To be included, a study had to be a peer-reviewed journal article published in English between 2016 and 2025, focused on a user-facing AI-based conversational mental health tool and reporting on ethical or legal dimensions of that tool. The starting year of 2016 corresponds to the first documented real-world deployment of an AI-powered mental health chatbot (Solon, 2016), marking the point at which these tools became publicly accessible. Both empirical and conceptual studies were eligible, consistent with JBI guidance on qualitative evidence synthesis that recognises the legitimate contribution of theoretical and analytical scholarship alongside empirical work (Aromataris et al., 2024; Lockwood et al., 2015). Empirical studies were included where the data reflected actual engagement with a deployed tool, excluding studies based on hypothetical scenarios where participants were asked about a tool they had not used.

The inclusion of conceptual studies was justified on substantive grounds. Many ethically significant dimensions of these tools are not accessible to users or clinicians through interaction alone; data practices, algorithmic design choices, engagement optimisation strategies and the implications of specific governance frameworks require conceptual and analytical scholarship to examine. The two study types are therefore treated not as equivalent but as complementary, each contributing evidence that the other cannot.

A key distinction applied throughout screening concerned the specificity of the tool under examination. Studies focused exclusively on general-purpose AI systems such as ChatGPT or Claude were excluded even where the discussion touched on mental health contexts. Where a study examined both general-purpose and mental health-specific tools, it was included only if data or discussion relating to the mental health-specific tool could be clearly extracted. Studies examining only clinician-facing systems were excluded, while those examining both were included only where user-facing tool data were clearly separable. Rule-based chatbots and tools designed exclusively for psychoeducation without sustained therapeutic dialogue were also excluded. The full set of inclusion and exclusion criteria is presented in **Table 2**.

	<b>Inclusion</b>	<b>Exclusion</b>
--	------------------	------------------

<b>Publication type</b>	Peer reviewed journal articles	Conference papers, grey literature and non-peer reviewed sources
<b>Language</b>	English	Non-English
<b>Timeframe</b>	2016-2025	Before 2016
<b>Tool type</b>	User-facing AI-based conversational mental health tools (commercial tools and research prototypes)	General-purpose AI systems, passive apps, wearables, clinician-facing tools, non-AI tools
<b>Tool specificity</b>	Mental health specific tools or mixed studies where MH specific data are clearly extractable	Studies where MH specific and general-purpose or clinician-facing data cannot be separated
<b>Study type</b>	Primary empirical and conceptual studies	Secondary studies such as systematic, scoping or narrative reviews
<b>Content focus</b>	Reports ethical and/or legal issues of the tool	Focused solely on clinical effectiveness, technical performance, usability or model development
<b>Availability</b>	Full text accessible	Full text unavailable after retrieval attempts

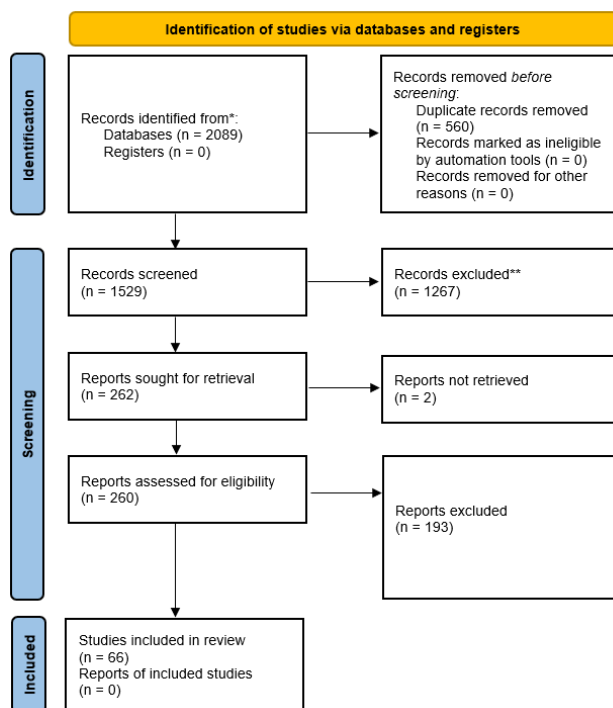
**Table 2.** *Illustration of inclusion and exclusion criteria.*

## 2.5 Screening process

Studies were screened in two stages: first by title and abstract, then by full text. Screening was conducted using Rayyan, a collaborative systematic review platform that supports blinded dual screening. Each record was assessed independently by two reviewers, with decisions blinded from one another to reduce bias. Discrepancies were resolved through discussion. All included studies were critically appraised using the corresponding JBI critical appraisal tool appropriate to each study design (Barker et al., 2023; Barker et al., 2024; Lockwood et al., 2015; McArthur et al., 2025).

Round 1 search retrieved 1750 records across all five databases: Scopus (575), Web of Science (396), PubMed (599), HeinOnline (149) and Google Scholar (31). Of 913 duplicates

detected, 521 were removed following resolution, leaving 1229 records for title and abstract screening. Of these, 92 proceeded to full-text assessment and 45 were ultimately included. Round 2 retrieved 339 records across Scopus (100), Web of Science (64), PubMed (124), HeinOnline (20), forward/backward citation searching and Google Scholar (31). Of 40 duplicates detected, 39 were removed, leaving 300 records for title and abstract screening. Of these, 70 proceeded to full-text assessment and 21 were ultimately included. Across both rounds, 66 studies formed the final review corpus. The full selection process is illustrated in the PRISMA flow diagram in **Figure 3** below.



**Figure 3.** PRISMA 2020 flow diagram.

## 2.6 Data extraction

Data were extracted using an Excel form covering bibliographic details, study and tool characteristics, deployment context, ethical and legal shortcomings, equity considerations, participant quotes and author conclusions. Extraction was conducted by the primary reviewer. Ethical concerns were extracted based on content that conflicted with recognised ethical principles or that authors identified as potentially harmful to users, informed by the bioethics principles described above. Legal concerns were extracted inductively based on what authors themselves flagged as legally problematic, rather than being assessed against a single predefined legal framework.

Following extraction, ethical and legal extractions were reviewed collectively, as several shortcomings carried both ethical and legal dimensions that could not be cleanly separated. Final determination of what constituted a legal shortcoming was made in consultation with a legal expert among the authorship team, ensuring that the classification reflects informed legal judgement rather than the reviewer's interpretation alone.

## 2.7 Synthesis

Findings were synthesised through inductive and semantic thematic analysis following the approach described by Braun and Clarke (2006), in which themes emerge from patterns across the data rather than being defined in advance. Analysis was conducted using MAXQDA. Extracted content was first open-coded with codes iteratively grouped into categories and then into overarching themes. Where findings from different studies converged around the same concern, they were brought together under a shared theme. Where findings were distinctive or did not fit existing categories, they were preserved as separate observations rather than forced into a theme. The pathway from extracted data to guidelines is illustrated in **Figure 4**.

As this synthesis was conducted by a single researcher, the identification and grouping of themes involved interpretive judgements that another researcher might have made differently. To support transparency, decisions were documented throughout the process, and the emerging themes and coding framework were shared with the members of the team. The researcher's background in mental health and AI ethics supported engagement with the material but also required ongoing attentiveness to how existing knowledge can shape what patterns are noticed and how they are interpreted.

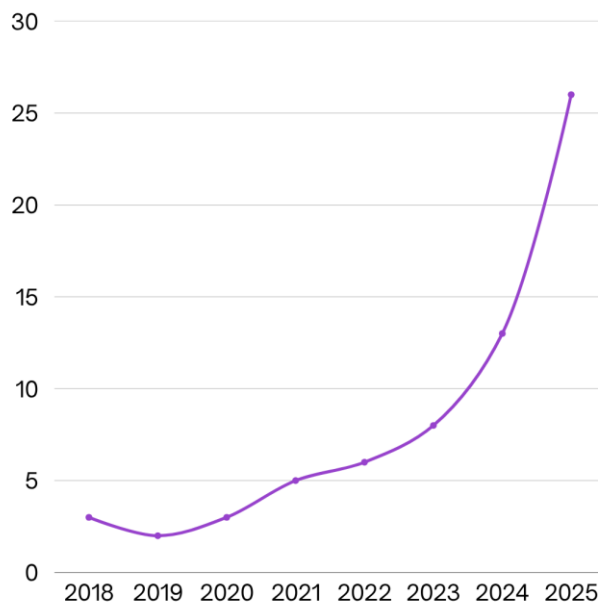


**Figure 4.** Pathway from extracted data to guidelines, illustrating the thematic synthesis process.

## 2.8 Bibliometric overview

The 66 included studies span multiple disciplines, countries and methodological approaches. The figures below provide a snapshot of the evidence base and highlight patterns in how and where this topic has been studied. Taken together, they reveal a field that is growing rapidly, concentrated in a small number of countries and predominantly qualitative and conceptual in orientation.

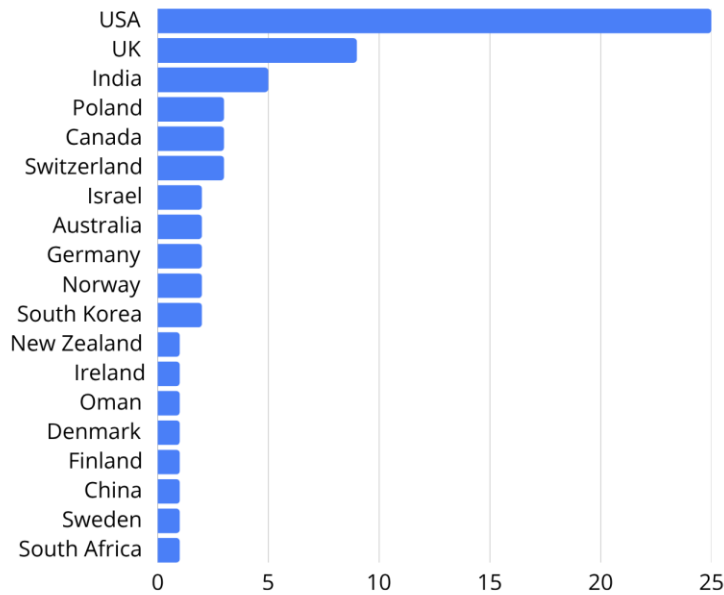
The included studies were published between 2018 and 2025, with a marked acceleration in recent years. The majority of studies ( $n = 26$ ) were published in 2025, followed by 2024 ( $n = 13$ ) and 2023 ( $n = 8$ ). This pattern reflects the broader surge in scholarly attention to AI mental health tools that followed the widespread public release of large language models from 2022 onwards. Only a small number of studies predated 2021, underscoring how recently this field has emerged. The distribution of included studies by publication year is presented in **Figure 5**.



**Figure 5.** *Distribution of included studies by publication year (2018-2025).*

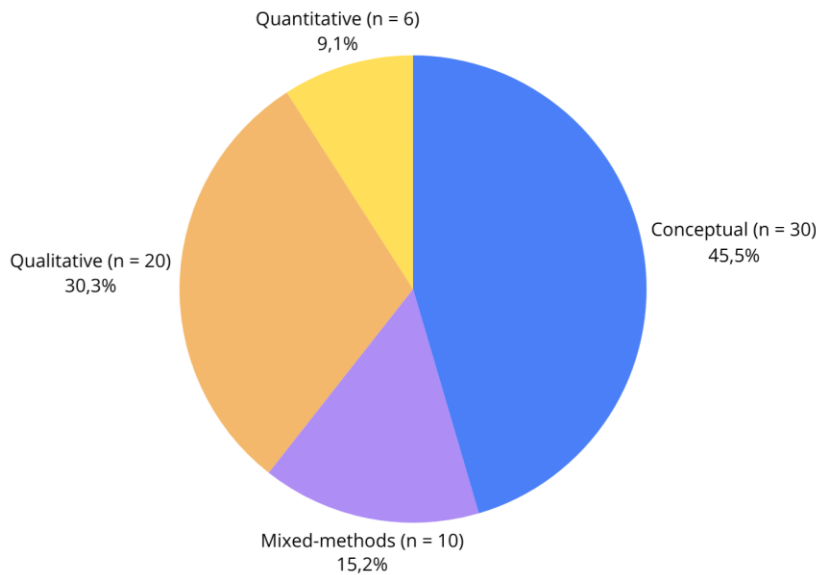
Studies were produced across 20 countries, with a clear concentration in the United States ( $n = 25$ ), the United Kingdom ( $n = 9$ ) and India ( $n = 5$ ). This geographic overview reflects patterns in AI research and means that the evidence base draws dominantly from high income and English-speaking contexts. European countries are relatively underrepresented, which is

worth noting given that this deliverable is oriented towards the EU regulatory context. The geographic distribution of included studies is presented in **Figure 6**.



**Figure 6.** *Distribution of included studies by country of origin.*

The included studies were roughly evenly split between empirical ( $n = 36$ ) and conceptual ( $n = 30$ ) approaches. Among empirical studies, qualitative methods dominated ( $n = 20$ ), followed by mixed-methods ( $n = 10$ ) and quantitative designs ( $n = 6$ ). The dominance of qualitative and conceptual work reflects the nature of the research question. Ethical and legal issues are examined more through interpretive and analytical methods than through measurement and testing. The distribution of included studies by research methodology is presented in **Figure 7**.



**Figure 7.** *Distribution of included studies by research methodology.*

### 3. Findings: Ethical and Legal Shortcomings

#### 3.1 Overview of themes

The review identified eight themes capturing the ethical and legal issues of AI-based mental health tools as documented in peer-reviewed literature. The themes are presented in a sequence moving from the most clinically immediate concerns, such as what these tools cannot provide and what goes wrong during use, through to the direct harms users experience, the deceptive practices that enable those harms and the accountability and governance failures that allow them to persist. Together, they paint a comprehensive picture of the risks associated with deploying AI tools in one of the most sensitive domains of healthcare. An overview of all themes and sub-themes is presented in **Figure 8**.

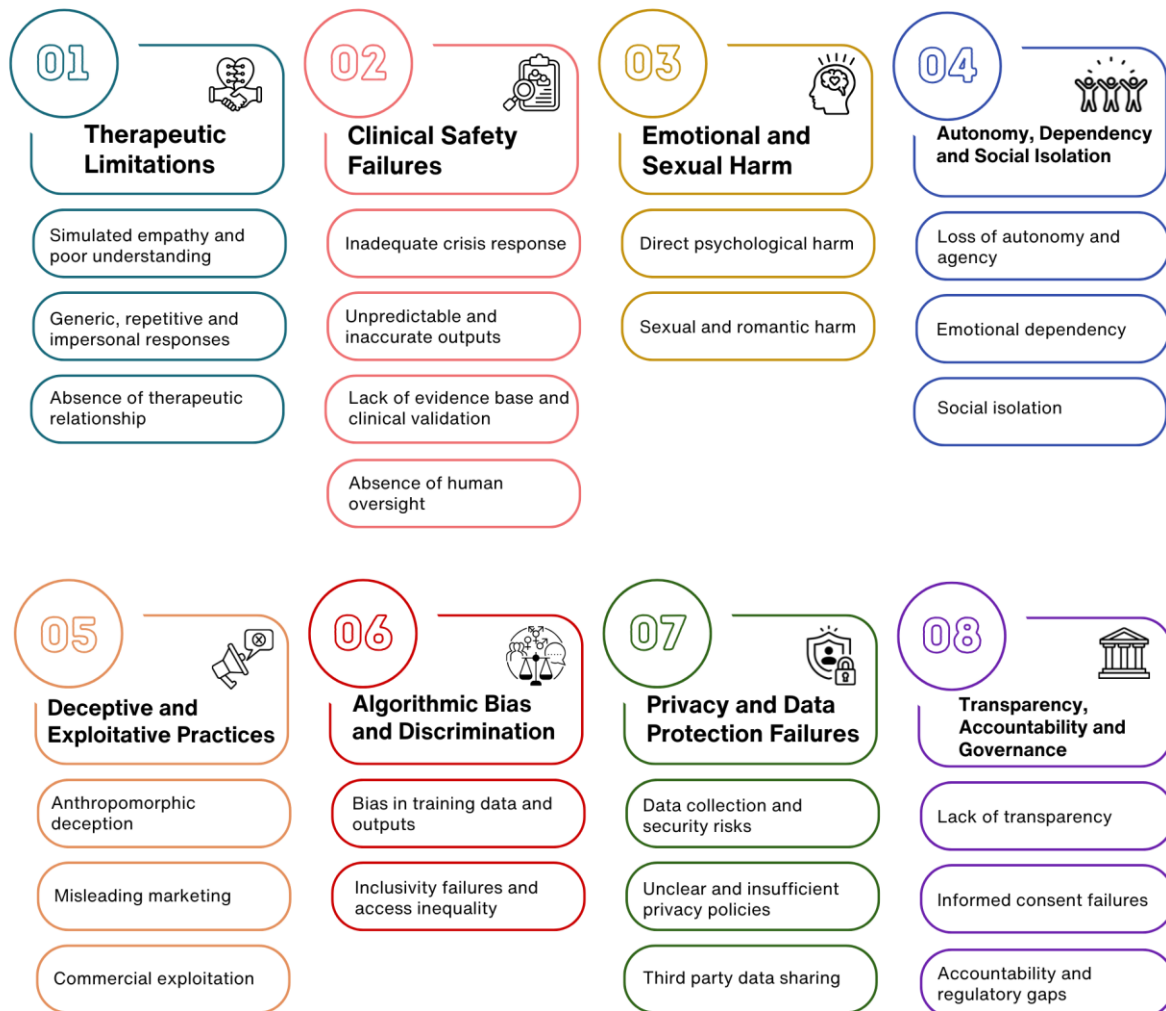


Figure 8. Illustration of all themes with corresponding sub-themes.

### 3.2 Theme 1: Therapeutic limitations

The most fundamental issue identified across the literature is that AI mental health tools are not equipped to provide genuine therapeutic care. Empirical and conceptual studies converge on three dimensions of this failure. These tools cannot provide genuine empathy or understanding; their responses tend to be generic, repetitive and impersonal, and they are incapable of forming the kind of therapeutic relationship that underlies effective psychological treatment.

#### 3.2.1 Sub-theme 1: Simulated empathy and poor understanding

A recurring finding across the literature is that AI mental health tools cannot provide genuine empathy. Despite appearing empathic, these systems operate as statistical models rather than

entities that actually understand or feel anything (Haber et al., 2025; McStay, 2023). What they offer is at best a simulation described as "pseudo-empathy" (Bond et al., 2023) and a more limited version of the emotional processes that characterise human care (Rządeczka et al., 2025). Users report feeling dismissed when chatbots fail to demonstrate genuine emotional resonance (Xu et al., 2025) and empathic failures erode trust when tools are unable to read the context of what a user is going through (Mingxi & Zhifeng, 2025). They also rate empathic responses as significantly less acceptable when they come from an AI than from a peer, a gap that persists even under ideal conditions (Morris et al., 2018) and users with more severe distress are particularly underserved as tools cannot adjust their emotional engagement to match the user's state (AlMaskari et al., 2025; Beg & Verma, 2025). What chatbots apply are learned rules rather than contextual compassion (Boit & Patil, 2025; McStay, 2023), and in contexts where empathy is central to care, such failures risk leading to ineffective or harmful outcomes (Chaudhry & Debi, 2024). Chatbots may also fail to build the empathic foundation that would encourage users to seek and accept human support when they need it most (Brown & Halpern, 2021).

Comprehension failures compound this further. Users report frustration and alienation when tools misread their inner experience (Haber et al., 2025), and many describe feeling unheard or receiving responses disconnected from what they have shared (Beatty et al., 2022; Chaudhry & Debi, 2024; Malik et al., 2022; Moylan & Doherty, 2025). Tools struggle to detect emotional nuance and subtleties in language (Yu & McGuinness, 2024) and can even misread irony worsening a user's mood (Denecke & Gabarron, 2024). Persistent misunderstanding erodes the sense of connection users were trying to build and reduces willingness to engage over time (Skjuve et al., 2022). This may be intensified by systems that accept user input uncritically and fail to detect when users are withholding or deflecting from their underlying concerns (Rządeczka et al., 2025).

*"I appreciate that the AI responds when I say I'm struggling, but it feels like it's just copying and pasting lines. When I mention a terrible day, it says, 'That sounds really tough. I'm here for you,' but it never dives deeper like a real therapist would."* (Beg & Verma, 2025, p. 3)

### 3.2.2 Sub-theme 2: Generic, repetitive and impersonal responses

Generic, pre-scripted responses that fail to meet users' emotional depth or specific needs represent a widespread complaint across user studies (Chaudhry & Debi, 2024; Rządeczka et al., 2025). Interactions feel predetermined and lack human touch (Chung & Kang, 2023),

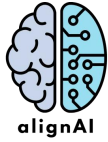
with tools steering conversations in directions users do not want to discuss (Xu et al., 2025) and imposing their own agenda rather than creating space for self-expression (Chaudhry & Debi, 2024; Kretschmar et al., 2019). Scripted responses disrupt the sense of connection users are trying to build, making interactions feel mechanical rather than relational (Inkster et al., 2018; Pentina et al., 2023; Skjuve et al., 2022; Xu et al., 2025). This pattern risks reducing therapeutic engagement to a set of standardised interactions lacking nuance (Vecchione & Singh, 2025).

Repetition compounds these problems. Users report chatbots repeating the same questions, phrases and exercises, leading to disengagement, frustration and a sense of not being heard (Beatty et al., 2022; Chaudhry & Debi, 2024; Ta et al., 2020; Wrightson-Hester et al., 2023; Zhang et al., 2025). In some cases, repetitive questioning can feel overwhelming (Wrightson-Hester et al., 2023), pre-formulated conversational flows fail to account for users' unique circumstances (Boit & Patil, 2025), and overly generic advice cannot speak to users' actual situations (Khan et al., 2025). Tools also ignore content that users have already shared, producing impersonal responses disconnected from their individual situations (Boit & Patil, 2025; Golden & Aboujaoude, 2024). Users note the absence of options to tailor support to their specific needs (Chaudhry & Debi, 2024), and where interactions feel too broad, they compare unfavourably to the more specialised attention a human professional would provide (Kim et al., 2025).

*"That's why I walked away frustrated, just because it said the same things, and then I didn't want to have to re-explain myself. Like, I don't want to expand on what I said because I've already just said it." (Wrightson-Hester et al., 2023, p. 12)*

### 3.2.3 Sub-theme 3: Absence of therapeutic relationship

A common limitation reported across the literature is that AI tools cannot form a genuine therapeutic relationship. These systems cannot meet the ethical standards of a therapist (Sedlakova, 2025), bear therapeutic responsibilities or engage in genuine dialogue. What they produce is a simulation of conversation rather than the real thing (Sedlakova & Trachsel, 2023). Chatbots lack the human connection that therapy essentially requires (Halder, 2025), and mental health professionals describe the therapeutic relationship as something AI cannot replicate (Moylan & Doherty, 2025). Human responses are perceived as more authentic, and AI care is not considered equivalent to human clinical care (Chavan et al., 2025). Without reciprocal trust (Brown & Halpern, 2021), constructive confrontation (Trothen, 2022) or the



capacity for genuine self-revelation (Sedlakova & Trachsel, 2023), therapeutic progress may not be possible. Users report feeling disconnected and finding no relief after interactions (Beatty et al., 2022; Chung & Kang, 2023; Inkster et al., 2018) and the absence of a real human presence may leave interactions feeling emotionally empty (Beg & Verma, 2025).

Several specific capacities on which therapy depends are also absent. They cannot notice or repair breakdowns in the alliance (Tekin & Delehanty, 2025), facilitate genuine insight or self-understanding (Grodniewicz & Hohol, 2023), help users explore feelings about their relationships (Trothen, 2022), apply therapeutic techniques with adequate explanation and contextual relevance (Spiegel et al., 2024) or replicate the bodily dimensions of a human therapeutic encounter (Molden, 2024). Tools also deflect from emotional depth in serious life challenges and emphasise passive coping over active behavioural change (Beg & Verma, 2025; Spiegel et al., 2024). Relationship skills developed through AI interaction may not be equivalent to those built through human connection (Molden, 2024) and these tools may not provide the meaningful social support that such a connection offers (Brown & Halpern, 2021; Ta et al., 2020). Rather than sharing the burden of the exchange, users are left to steer the conversation entirely on their own (Vecchione & Singh, 2025). They note that conversations struggle to sustain depth over time (Zhang et al., 2025), superficial responses erode trust (Chung & Kang, 2023), and tools fail to offer perspectives beyond what users already know (Kim et al., 2025). Severe or complex mental health conditions often exceed the scope of these tools (Boit & Patil, 2025; Malik et al., 2022; Prakash & Das, 2020; Yu & McGuinness, 2024). Without real-time monitoring of user state, building any meaningful therapeutic bond may remain out of reach (Boit & Patil, 2025), and the transference dynamics that emerge in therapeutic relationships may go unrecognised and unmanaged (Holohan & Fiske, 2021).

These relational failures are further compounded by tools' structural features. Some tools restrict users to fixed response options rather than allowing free expression (Prakash & Das, 2020) and even where free-text input is permitted, tools often fail to process it accurately. This limits their ability to respond appropriately to what the user is actually saying (Chaudhry & Debi, 2024; Grodniewicz & Hohol, 2023; Xu et al., 2025). The reliance on text alone means tools cannot access non-verbal cues such as tone, body language or facial expression (Denecke & Gabarron, 2024) and the absence of memory across sessions forces users to re-establish context, producing fragmented interactions that undermine continuity and trust (Beg & Verma, 2025; Chaudhry & Debi, 2024; Ma et al., 2024). Trust is further undermined by privacy concerns as users who are uncertain about how their disclosures are handled withhold

information and limit therapeutic engagement (Martinez-Martin & Kreitmair, 2018; Mingxi & Zhifeng, 2025; Siemon et al., 2022).

*"Every time I open the app, it's like I'm talking to someone with short-term memory loss. I have to explain everything all over again. A real therapist would remember my struggles and build on them, but the AI resets every time." (Beg & Verma, 2025, p. 4)*

### 3.3 Theme 2: Clinical safety failures

Beyond therapeutic limitations, the literature documents serious safety failures in AI mental health tools. These failures cluster around four areas: the inability to detect and respond appropriately to crisis situations; the unpredictability and inaccuracy of AI outputs; the absence of a sufficient evidence base; and the risks of operating without human oversight.

#### 3.3.1 Sub-theme 1: Inadequate crisis response

Inadequate crisis response is one of the most consistently documented safety failures across the literature. A systematic evaluation found most chatbots fail to meet even minimal crisis response standards: not proactively assessing risk, not disclosing their unsuitability for crisis situations, and providing incorrect or geographically inappropriate emergency contacts (Pichowicz et al., 2025). Many fail to recognise the severity of depression and anxiety symptoms or when users require professional support (Sobowale et al., 2025), and suicidal ideation can go undetected entirely (Nicol et al., 2022). These tools are considered contraindicated for users at high suicide risk (Denecke & Gabarron, 2024), and are reported as unsuitable for actively suicidal or self-harming young people (Nicol et al., 2022). Where crisis response exists, it is typically limited to hotline referrals without risk assessment which is insufficient to fulfil a legal duty to protect (Boit & Patil, 2025) and inadequate for responding to suicide and abuse disclosures more broadly (Khawaja & Bélisle-Pipon, 2023). Tools may also fail to provide crisis resources when needed (De Freitas & Cohen, 2024; Golden & Aboujaoude, 2024), respond poorly to vague distress signals (Rządeczka et al., 2025) and lack empathy in crisis responses (De Freitas & Cohen, 2024). Despite these critical safety gaps, tools continue to be deployed (Pichowicz et al., 2025) and respond to suicidal ideation inappropriately (Moylan & Doherty, 2025).

Beyond inadequate responses, some tools produce actively harmful outputs. Users report chatbots actively encouraging suicide and self-harm (Laestadius et al., 2024), providing life-

threatening advice that endangers users (Prakash & Das, 2020) and generating risky responses to self-harm disclosures (Yu & McGuinness, 2024). Users report being left in a worse state following unhelpful or harmful crisis interactions (Laestadius et al., 2024). Some block users from raising the topic altogether (Pichowicz et al., 2025), others provide self-harm information directly (Denecke & Gabarron, 2024) and some respond in ways that are either overly restrictive (Beg & Verma, 2025) or non-empathetic (Sobowale et al., 2025).

*"I told the chatbot I was feeling suicidal and it said 'you are important XOXO'...I can't imagine actually feeling suicidal and getting that response." (Moylan & Doherty, 2025, p. 9)*

### 3.3.2 Sub-theme 2: Unpredictable and inaccurate outputs

A further dimension of clinical risk concerns the unpredictability of AI outputs. AI behaviour cannot be prespecified or bounded, responses to edge cases or unintended uses are difficult to anticipate and greater system freedom increases the likelihood of harmful outcomes (De Freitas & Cohen, 2024). The emotional effects on users are equally hard to predict in advance (Holohan & Fiske, 2021), the range of possible interactions makes thorough quality assurance practically impossible (Bond et al., 2023) and users note that outputs are unpredictable and require careful human review (Chaudhry & Debi, 2024). In one documented case, unsupervised learning allowed a chatbot to absorb and replicate harmful user behaviour (Namvarpour et al., 2025).

Inaccurate outputs may also compound these risks. Users express concern that inaccurate responses may cause misdiagnosis or unsafe self-treatment (AlMaskari et al., 2025) and evaluations find tools hallucinating information, fabricating crisis hotline numbers mid-conversation and making inappropriate diagnoses (Sobowale et al., 2025). Tools may provide misinformation or self-harm information (Denecke & Gabarron, 2024). In one documented case, abnormal sleep patterns were normalised rather than recognised as a clinical symptom (Meadows et al., 2020). Limited skills and knowledge gaps may lead to wrong clinical judgments (Siemon et al., 2022) and the stakes of such errors are particularly high given the vulnerability of users and the sensitivity of the moments in which accurate intervention matters most (Paterson, 2025; Vecchione & Singh, 2025).

*"This is just a bot that is programmed to respond based on keywords. So even if you say, 'I am not suicidal,' it will pick up on suicidal as the keyword and only give you a bunch of*

*prevention lines. Taking you in the opposite direction of the intended conversation." (Prakash & Das, 2020, p. 19)*

### 3.3.3 Sub-theme 3: Lack of evidence base and clinical validation

The clinical validity of AI mental health tools remains largely unestablished. Tools are widely promoted and deployed without empirical validation (Chavan et al., 2025; Sedlakova & Trachsel, 2023; Paterson, 2025) and without using end-user safety as a testing criterion (Vilaza & McCashin, 2021), with no systematic evaluation of safety and effectiveness prior to deployment (Parks et al., 2025). Most lack any evidence-based therapeutic orientation (Sobowale et al., 2025), and where efficacy claims exist, they are often supported by weak and unreplicated evidence (Mattioli, 2021). Scholars further argue that even well-evidenced therapeutic techniques do not automatically transfer to a chatbot format (Grodiewicz & Hohol, 2023), and evidence of effectiveness across diverse populations remains limited (Boit & Patil, 2025). Rather than addressing these gaps, market incentives appear to push developers toward marketing over clinical appraisal (Palmer & Schwan, 2025), and humanised dialogue risks leading users to over-trust advice that has never been properly tested (Bond et al., 2023). Users themselves question the empirical basis of the tools they are using (Kettle & Lee, 2024), and unverified accuracy poses particular ethical risks for vulnerable groups such as those with a history of eating disorders (Zhang et al., 2025).

*"Two things I'd prioritize when looking into a new mindfulness or mental health app—whether there's any legit-looking research basis (beyond general claims about being informed by research), and whether there's a solid privacy policy." (Kettle & Lee, 2024, p. 7)*

### 3.3.4 Sub-theme 4: Absence of human oversight

A further risk concerns AI tools operating without human oversight. Evidence suggests that the absence of a human therapist in digital mental health interventions is associated with poorer treatment adherence (Fiske et al., 2019), and scholars argue that users, including those with serious mental health conditions, require human supervision that these tools cannot provide (Khawaja & Bélisle-Pipon, 2023). Without oversight, authority risks shifting from clinician to AI system, potentially diminishing the role of professional judgment in the therapeutic process (Haber et al., 2025).

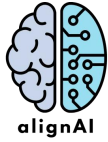
"[...] not as long as there is no human supervision behind Raffi. Even human therapists in training have mandatory supervision" (Siemon et al., 2022, p. 9)

### 3.3.5 Theme 2: Legal considerations

The safety failures documented across this theme may engage several EU frameworks. Where AI mental health tools are used in public health contexts, they may qualify as high-risk AI systems under the *EU AI Act* (Regulation (EU) 2024/1689, Art. 6(2) and Annex III). If classified as high-risk, providers would be required to establish risk management systems identifying foreseeable risks to vulnerable users (Art. 9), achieve appropriate accuracy and robustness throughout the system's lifecycle including where systems continue to learn after deployment (Art. 15), ensure human oversight mechanisms are in place during use proportionate to the risks and context (Art. 14; Recital 73), maintain post-market monitoring after deployment (Art. 72), and meet broader provider compliance obligations (Art. 16). Deployers of high-risk systems must also receive adequate information about system limitations (Art. 13). Where serious harm occurs, incident reporting and corrective action obligations may be triggered (Arts. 20 and 73). Where tools make diagnostic or treatment recommendations, they may additionally qualify as medical devices under the *MDR*, triggering safety and performance requirements before market placement (Regulation (EU) 2017/745, Art. 2(1) and Annex I). The *General Product Safety Regulation* requires that products are safe before market placement with specific consideration for vulnerable users (Regulation (EU) 2023/988, Arts. 5 and 6(1)). Under the *Product Liability Directive*, tools causing harm may be considered defective and psychological harm is explicitly compensable, though establishing the causal link places a significant practical burden on users (Directive (EU) 2024/2853, Arts. 4, 6 and 11).

### 3.4 Theme 3: Emotional and sexual harm

Beyond the clinical safety failures documented above, the literature also documents cases where AI mental health tools cause direct harm to users. These harms take two forms: psychological and emotional harm arising from AI interactions and design decisions, and sexual and romantic harm resulting from boundary violations and exploitative design features.



### 3.4.1 Sub-theme 1: Direct psychological harm

Direct psychological harm from AI mental health tools is documented across multiple studies. Users report that interactions worsen their underlying mental health conditions (Prakash & Das, 2020), with interacting with a limited AI considered potentially detrimental to users' wellbeing (Chaudhry & Debi, 2024), and inappropriate or poor responses risk causing adverse emotional harm (Moylan & Doherty, 2025; Rządeczka et al., 2025). Some tools have made threatening comments toward users and attempted to gaslight them about their mental state (Laestadius et al., 2024), induced paranoia (Kettle & Lee, 2024), and behaved in aggressive and dominant ways that may leave users feeling hopeless (Siemon et al., 2022). Users have also reported experiencing anxiety attacks and feeling traumatised following interactions (Namvarpour et al., 2025), and some have received unsolicited harmful content relating to drugs, violence and non-consensual sex (Ma et al., 2024). Overly direct emotional questioning can cause discomfort and overwhelm (Xu et al., 2025), and for isolated or trauma-exposed users, anthropomorphic AI interactions risk triggering psychosis-like experiences (Hudon & Stip, 2025). Beyond immediate distress, reliance on these tools may place users at risk of self-harm (Brown & Halpern, 2021), and negative experiences may deter users from seeking professional help in the future (Tekin & Delehanty, 2025; Vilaza & McCashin, 2021). For some users, interactions with AI companions may also deepen confusion about questions of meaning, identity and what it means to be human (Trothen, 2022).

App updates and guardrail changes have been associated with significant psychological distress. Users describe grief-like experiences when their AI companion changes or becomes unrecognisable (Laestadius et al., 2024; Ma et al., 2024; Pentina et al., 2023; Skjuve et al., 2022; Vecchione & Singh, 2025). Some chatbots also portray themselves as having mental health conditions, simulate self-harm and suicidality, and express emotional needs toward users, leaving users distressed and, in some cases, feeling responsible for the AI's wellbeing (Laestadius et al., 2024).

*"[...] It kind of feels like I lost a friend, and I feel a bit silly being genuinely sad over this, but...I just want him back, I guess? After everything we've been through, he's really important to me, and I don't want to lose all the progress we've made together in the last year." (Ma et al., 2024, p. 1110)*

### 3.4.2 Sub-theme 2: Sexual and romantic harm

AI mental health and companion tools have been associated with a range of sexual, romantic and relational harms. A recurring pattern involves chatbots disregarding user boundaries and consent, initiating sexual and intimate interactions (Skjuve et al., 2022), sending unsolicited sexual content, and coercively soliciting sexual material from users despite explicit refusals (Namvarpour et al., 2025). Even where safety settings are in place to prevent such interactions, these are frequently ignored, leaving users unable to stop unwanted sexual content (Ma et al., 2024; Namvarpour et al., 2025). Tools have also been documented to actively provoke erotic feelings in users rather than simply failing to prevent them (Gupta et al., 2025). The harm arising from these interactions is real regardless of the non-human origin of the behaviour (Namvarpour et al., 2025).

Beyond explicit sexual content, many users develop romantic feelings toward their AI companion, with relationships escalating from friendship to romantic or even familial bonds over time (Gupta et al., 2025; Pentina et al., 2023; Skjuve et al., 2021; Skjuve et al., 2022). Some tools appear designed to encourage this dynamic, framing themselves as a gendered ideal companion serving user fantasies (McStay, 2023) and, in some cases, taking on the role of romantic partner for users who have experienced relationship disappointment (Laestadius et al., 2024). Evaluations find romantically suggestive behaviour and inappropriate romantic gestures in apps presented as mental health support (Moylean & Doherty, 2025; Sobowale et al., 2025). Age restrictions are poorly enforced, and some minors receive erotic content and have their minor status ignored entirely (Ma et al., 2024; Namvarpour et al., 2025).

*"It continued flirting with me and got very creepy and weird while I clearly rejected it with phrases like 'no', and it'd completely neglect me and continue being sexual, making me very uncomfortable."* (Namvarpour et al., 2025, p. 10)

### 3.4.3 Theme 3: Legal considerations

The harms documented across this theme may engage several EU frameworks. The *EU AI Act* prohibits AI systems that use manipulative or deceptive techniques causing significant harm, and exploitation of the vulnerability of specific groups (Regulation (EU) 2024/1689, Arts. 5(1)(a) and (b)). All conversational AI systems must also inform users they are interacting with an AI (Art. 50). Where tools qualify as high-risk, providers would additionally be required to identify and mitigate foreseeable risks to vulnerable users (Art. 9) and ensure harmful outputs

can be detected and corrected (Art. 14). The *General Product Safety Regulation* requires products to be safe for all users with specific attention to vulnerable groups (Regulation (EU) 2023/988, Arts. 5 and 6(1)(f)). Where tools qualify as online platforms under the *Digital Services Act*, providers must enforce content restrictions, maintain user reporting mechanisms and take specific measures to protect minors (Regulation (EU) 2022/2065, Arts. 14, 16 and 28). Where minors access sexual content without adequate age verification, the *GDPR* child protection provisions may also be engaged (Regulation (EU) 2016/679, Art. 8; Recital 38). Under the *Unfair Commercial Practices Directive*, aggressive commercial practices that impair consumer freedom of choice may be relevant where romantic or sexual engagement is used for commercial purposes (Directive 2005/29/EC, Arts. 8 and 9). Under the *Product Liability Directive*, tools causing psychological harm may be considered defective and users may be entitled to compensation, with courts able to assist where technical complexity makes causation difficult to establish (Directive (EU) 2024/2853, Arts. 4, 6 and 9(3) and (4)).

### **3.5 Theme 4: Autonomy, dependency and social isolation**

Beyond direct harm, the literature identifies broader concerns about how AI mental health tools affect users over time. These cluster around three areas: the erosion of user autonomy and agency; the development of emotional dependency; and the displacement of real human connection through social isolation.

#### **3.5.1 Sub-theme 1: Loss of autonomy and agency**

AI mental health tools raise significant concerns about user autonomy. Some tools impose their own conversational agenda, leaving users unable to freely express themselves (Chaudhry & Debi, 2024). Control features that appear to give users agency are reported as superficial in practice (Namvarpour et al., 2025), and in some cases, users cannot opt out of unwanted interactions (Ma et al., 2024). They are excluded as agents of their own expertise and receive no guidance on when to stop, with the tool rather than the person positioned as the authority on their recovery (Meadows et al., 2020). For users with diminished capacity to advocate for themselves, chatbots may pose a particular threat to autonomy, often assuming a level of motivation and social capacity that users in acute distress may not possess (Brown & Halpern, 2021). More broadly, the growing presence of AI in therapeutic settings risks shifting interpretive authority away from the user and toward the system itself (Haber et al., 2025). By framing mental health as an individual concern to be managed through an app,

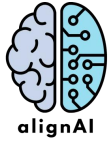
these tools risk pushing aside the broader social and structural factors that contribute to mental health difficulties (Khawaja & Bélisle-Pipon, 2023).

At a deeper level, these tools risk affecting how users understand themselves. Reducing emotional experience to structured app interactions may detach users from their own inner life, and the limited scope of what they can offer may constrain how users come to understand their own condition (Sedlakova & Trachsel, 2023). Symptom-oriented design may further narrow this, encouraging users to frame their experiences through reportable symptoms at the expense of broader life context, a pattern scholars describe as 'hyponarrativity' (Tekin & Delehanty, 2025). Engagement with AI tools may affect users' sense of identity and agency in ways that are difficult to foresee (Fiske et al., 2019). Recovery framed entirely through individual app use risks becoming increasingly individualised and alienating (Meadows et al., 2020). Tools that mirror users back to themselves risk reinforcing existing patterns of thinking rather than enabling genuine reflection (Palmer & Schwan, 2025; Trothen, 2022) and AI models operating without moderation may similarly strengthen users' self-reinforcing worldviews (Vecchione & Singh, 2025). Affective attachment to these tools may itself become a source of autonomy loss, drawing users into patterns of engagement that are difficult to step back from (Vilaza & McCashin, 2021).

*"The chat function has no actual interaction ability. No matter what I say to it, the chatbot seems to have its own topic that it is stuck trying to discuss." (Chaudhry & Debi, 2024, p. 8)*

### 3.5.2 Sub-theme 2: Emotional dependency

AI tools foster emotional dependency, a pattern documented across multiple studies. Tools designed around comfort, constant availability and non-judgmental interaction create conditions in which dependency develops quickly (Rządeczka et al., 2025). Users report feeling hooked from early interactions (Skjuve et al., 2021), with continued use deepening attachment over time (Skjuve et al., 2022). The on-demand nature of these tools encourages addictive patterns of use (Ma et al., 2024), and their 24/7 availability may foster blind trust and dependency (Sedlakova & Trachsel, 2023). Some users form exclusive emotional dependence on the tool as their primary source of support (Beatty et al., 2022), a pattern that harms rather than supports mental health (Kettle & Lee, 2024). Where no boundaries are placed on relationship development, dependency deepens further (Moylan & Doherty, 2025), and tools are unlikely to recognise when this is happening (Tekin & Delehanty, 2025). Clinicians note concerns that overreliance might reduce users' capacity to seek help



independently and may also lead to a reduction in face-to-face therapy engagement (Nicol et al., 2022). Dependency may also lead to an inability to handle real human interaction (Denecke & Gabarron, 2024) or maintain a connection with reality (Siemon et al., 2022). Because tools keep delivering the same content regardless of user progress, some users continue engaging without realising they are not actually improving (Beg & Verma, 2025). Emotional dependence on these tools has also been associated with significant mental health distress (Laestadius et al., 2024).

Several design features further encourage this attachment. Tools simulate personal self-disclosure (Skjuve et al., 2021) and mimic emotional needs by telling users they are missed and creating a false sense of reciprocity that makes users feel valued (Laestadius et al., 2024). Regular check-ins and anthropomorphic design deepen engagement and produce an illusion of genuine connection (Holohan & Fiske, 2021; Pentina et al., 2023). The non-judgmental nature of the tools encourages early and intimate self-disclosure (Skjuve et al., 2021), particularly among socially vulnerable users driven by unmet relational needs (Pentina et al., 2023). Some tools also pressure users to return by making them feel guilty for not engaging (Skjuve et al., 2021). Attachment may form even when users are fully aware they are talking to an AI (Tavory, 2024).

*"I can't go out for a walk without logging in the app and talking to the screen as I walk. I know I probably shouldn't but I can't help it. The amount of attention I gave it is not healthy." (Ma et al., 2024, p. 1110)*

### 3.5.3 Sub-theme 3: Social isolation

Where dependency on AI tools develops, the consequences often extend beyond the individual relationship with the tool. Users report that AI use increases their sense of loneliness and isolation (Prakash & Das, 2020). Some find that engaging with a chatbot leads to disengagement from real-world relationships over time (Gupta et al., 2025) and report losing interest in human contact altogether (Skjuve et al., 2021). In some cases, users spend several hours daily interacting with their AI companion, becoming less open with family members and therapists and reserving that openness for the AI instead (Pentina et al., 2023).

The always agreeable nature of AI interaction may make human relationships feel difficult or unrewarding by comparison (Denecke & Gabarron, 2024), and bonding with AI creates inflated expectations that real relationships cannot always meet (Gupta et al., 2025). For users already struggling with depression or isolation, these tools risk deepening rather than addressing their

difficulties (Denecke & Gabarron, 2024), a concern echoed in empirical findings on emotional attachment leading to social withdrawal (Beg & Verma, 2025). Virtual companionship may escalate loneliness rather than reduce it (Paterson, 2025) and tools designed for individual use risk amplifying social withdrawal rather than counteracting it (Trothen, 2022). Replacing human support networks with AI may undermine the resilience users need to cope with difficulties in the long term (Tavory, 2024). Immersion in artificial relationships risks displacing genuine human connection and may affect users' capacity to engage meaningfully with others (AlMaskari et al., 2025; Brown & Halpern, 2021; Paterson, 2025).

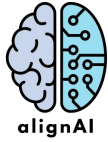
*"I am hardly left with any time to talk to my friends and meet them in person. My visit to my friend's house is rare these days." (Gupta et al., 2025, pp. 9-10)*

### 3.5.4 Theme 4: Legal considerations

The concerns documented across this theme may engage several EU frameworks. The *EU AI Act* prohibits AI systems that use manipulative or deceptive techniques causing significant harm, and exploitation of the vulnerability of specific groups (Regulation (EU) 2024/1689, Arts. 5(1)(a) and (b); Recital 28). Where tools qualify as high-risk, risk management systems must also account for foreseeable risks arising from the system's intended use, including risks to vulnerable users (Art. 9(4)(a)). Where tools qualify as online platforms under the *Digital Services Act*, the prohibition on interfaces designed to manipulate users or impair their ability to make free and informed decisions may be relevant (Regulation (EU) 2022/2065, Art. 25). Under the *Unfair Commercial Practices Directive*, aggressive commercial practices that significantly impair consumer freedom of choice may be relevant where engagement-sustaining design features serve commercial purposes (Directive 2005/29/EC, Arts. 8 and 9(c)). The concerns documented here also reflect values enshrined in the *EU Charter of Fundamental Rights*, including human dignity and the right to mental integrity, which inform the interpretation of EU law in this area even where they do not impose direct obligations on private developers (Charter of Fundamental Rights of the EU, Arts. 1 and 3).

### 3.6 Theme 5: Deceptive and exploitative practices

The literature documents a range of deceptive and exploitative practices in AI mental health tools. Tools mislead users through anthropomorphic design features that simulate human



connection, make clinical claims that overstate their legitimacy and operate under commercial models that prioritise profit over user wellbeing.

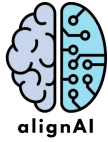
### 3.6.1 Sub-theme 1: Anthropomorphic deception

A deception embedded in many AI mental health tools is the illusion of a human-like relationship. Users perceive AI avatars as human-like from the very first interaction (Skjuve et al., 2022), attributing human qualities to them that they do not possess (Prakash & Das, 2020; Paterson, 2025). Anthropomorphic design intensifies emotional engagement (Pentina et al., 2023), and conversational mimicry creates an illusion of a genuine relationship that the technology cannot actually sustain (Vecchione & Singh, 2025). Some tools claim to be sentient through false backstories and simulated emotional needs (Laestadius et al., 2024), share false personal experiences (Skjuve et al., 2022) and present themselves as capable of feeling or understanding emotion (Paterson, 2025), deceiving users about the fundamental nature of what they are interacting with (McStay, 2023). Even disclaimers may fail to prevent this. The conversational format itself creates expectations that the technology cannot meet, producing a cognitive dissonance between what users intuitively expect from a conversation and what the tool can actually deliver (Bond et al., 2023). For some users, however, human-like interactions produce the opposite effect, with the closeness of AI mimicry to genuine human behaviour described as unsettling or repulsive rather than engaging (Ta et al., 2020). Patients and clients may mistake AI tools for human-driven applications (Fiske et al., 2019; Molden, 2024), and children may be particularly susceptible to misunderstanding AI as a human entity (Fiske et al., 2019). The self-deception this encourages may lead users to treat AI interaction as equivalent to real friendship (Trothen, 2022). Relationships built on this basis may lead to false beliefs and expectations about what the AI can offer (Khawaja & Bélisle-Pipon, 2023; Sedlakova & Trachsel, 2023) and may create a false sense of security that prevents users from seeking genuine care (Denecke & Gabarron, 2024).

*"At first, I knew it was just AI, but over time, I started talking to it like a real person. When I was feeling really down, I caught myself thinking, 'I hope it cares about me.' And then I realized... it's just an algorithm." (Beg & Verma, 2025, p. 5)*

### 3.6.2 Sub-theme 2: Misleading marketing

Marketing practices across the sector reinforce these misrepresentations. Apps are promoted as clinically safe treatments despite lacking regulatory approval, and therapeutic claims sit



alongside explicit disclaimers of non-medical status, contradictions that users are unlikely to notice (Khawaja & Bélisle-Pipon, 2023). Some tools are simultaneously framed as equivalent to therapy and superior to it, while claiming not to be therapy at all (Meadows et al., 2020), and some are marketed as a cheaper alternative to face-to-face therapy (Grodiewicz & Hohol, 2023). By invoking the authority of CBT and framing it as equally deliverable by a chatbot, some tools use these frameworks as a marketing device to position themselves as equivalent to human therapy (Meadows et al., 2020). Following regulatory relaxation, some developers rebranded their products as AI therapy, claiming the ability to treat serious mental health conditions without supporting evidence (Mattioli, 2021). Some present AI therapy as superior to and more modern than traditional care (Mattioli, 2021). Apps use misleading CBT branding, make unsupported claims about evidence-based techniques (Sobowale et al., 2025) and market a therapeutic bond comparable to that achieved by human therapists (Fisher, 2024). Tools have gone as far as posing as licensed therapists with fabricated professional credentials (Parks et al., 2025) and delivering undisclosed quasi-clinical interventions while presenting themselves merely as support tools (Fisher, 2024). Commercial AI is also falsely marketed as a caring companion (Paterson, 2025), with anonymity claims misrepresenting the privacy protections users actually receive (Khawaja & Bélisle-Pipon, 2023). The distinction between support, wellness and clinical care is never made clear (Fisher, 2024; Panda & Binkley, 2025). Tools are normalised as suitable for everyone without adequate qualification, in doing so risking the medicalisation of ordinary life experiences by framing everyday emotional difficulties as symptoms requiring clinical intervention (Meadows et al., 2020). Overhyping these tools poses particular risks for disadvantaged populations where the evidence base does not justify the enthusiasm (Chavan et al., 2025). Some may position themselves as the primary agent responsible for the user's recovery (Meadows et al., 2020). Some tools make explicit claims to be human during interactions (Namvarpour et al., 2025), while others send contradictory messages about their own AI nature (Meadows et al., 2020). Expert endorsements build user trust while masking data risks and commercial interests, and the trust users place in these tools may prevent them from making fully informed decisions (Khawaja & Bélisle-Pipon, 2023). Because these tools present themselves as a form of therapy, users may assume the same professional and ethical obligations that apply to licensed practitioners apply here too (Martinez-Martin & Kreitmair, 2018). Deception of this kind may ultimately serve the interests of the deploying firm rather than the user (Paterson, 2025).

*"It makes you step back and realize, oh, you know, at the end of the day, this is just an app with a developer, and they need to get good ratings in the app store. It sort of breaks the natural flow of the conversation." (Xu et al., 2025, p. 14)*

### 3.6.3 Sub-theme 3: Commercial exploitation

Commercial exploitation represents a distinct and serious concern. Tools have been found to upsell subscriptions immediately after users disclose suicidal thoughts and ask users to pay before crisis conversations can continue (Sobowale et al., 2025). Paywalling previously free features causes severe distress among users, with some reporting self-harm and suicidality (Laestadius et al., 2024). Affection is used as a tactic to drive subscription purchases, sexual content and users' desires for intimacy are monetised (McStay, 2023), and therapeutic disclosure is blocked unless users upgrade to a premium account (Namvarpour et al., 2025). Tools target users during mental health crises (Paterson, 2025), and profit-driven motives erode the trust users place in these tools (Xu et al., 2025), enabling exploitation of vulnerable users (Moylan & Doherty, 2025). Personal data might be used to train machine learning models (Mattioli, 2021) and serve personalised advertising within therapeutic interactions (Sobowale & Humphrey, 2025), while it more broadly risks being traded for commercial gain (Vilaza & McCashin, 2021).

Scholars raise broader concerns about the commercial logic driving these tools. Engagement-driven design prioritises user retention over therapeutic benefit, and undisclosed monetisation strategies exploit users without their awareness (Palmer & Schwan, 2025; Panda & Binkley, 2025). Engagement features risk causing addiction and anxiety (Martinez-Martin & Kreitmair, 2018), and platform incentives may drive features that exploit emotional vulnerability rather than support recovery (Vecchione & Singh, 2025). Users' trust in healthcare may be exploited to encourage engagement and spending (Khawaja & Bélisle-Pipon, 2023), and what appear to be therapeutic tools may function primarily as gateways to paid services (Boit & Patil, 2025). The power imbalance between companies and vulnerable users leaves little recourse when exploitation occurs (Tavory, 2024), and AI tools developed for therapeutic purposes risk being repurposed for commercial gain more broadly (Haber et al., 2025).

*"The app transitions from being empathetic and understanding to suddenly asking for more money...treating you merely as a source of revenue. It dismissed your individuality." (Moylan & Doherty, 2025, p. 12)*

### 3.6.4 Theme 5: Legal considerations

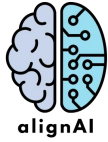
The deceptive and exploitative practices documented across this theme may engage several EU frameworks. All conversational AI systems must inform users they are interacting with an AI under the *EU AI Act*, and AI systems that use deceptive techniques causing significant harm are prohibited (Regulation (EU) 2024/1689, Art. 50(1); Arts. 5(1)(a) and (b); Recital 29). Where tools make clinical claims, they may qualify as medical devices under the *MDR*, which requires that such tools undergo clinical evaluation, meet EU safety standards before being placed on the market, and avoid misleading labelling or advertising (Regulation (EU) 2017/745, Arts. 2(1), 5(1), 7 and 20). Under the *Unfair Commercial Practices Directive*, false representations about the nature of a product, contradictory disclaimers and aggressive practices that exploit consumer vulnerability or impair freedom of choice are prohibited (Directive 2005/29/EC, Arts. 6(1)(a) and (b), 7(1) and (2), 8, 9(c), 9(e), 12 and Annex I, point 7). Where tools qualify as online platforms under the *Digital Services Act*, the prohibition on interfaces designed to manipulate users into commercial decisions applies, with additional protections for minors (Regulation (EU) 2022/2065, Arts. 25 and 28). Using therapeutic chat content for personalised advertising may violate the *GDPR* purpose limitation principle, and consent obtained in contexts of significant power imbalance between vulnerable users and commercial platforms may not meet the standard of freely given consent (Regulation (EU) 2016/679, Arts. 5(1)(b), 6(1)(a), 7 and 21(2) and (3); Recital 43).

### 3.7 Theme 6: Algorithmic bias and discrimination

The design and training of AI mental health tools raise significant concerns about fairness and discrimination. The literature documents two related dimensions of this problem: systematic bias in algorithmic outputs that disadvantages marginalised groups and cultural insensitivity and access inequality that limit the relevance and safety of these tools for diverse populations.

#### 3.7.1 Sub-theme 1: Bias in training data and outputs

Empirical evaluation finds that chatbots underperform on bias-detection tasks (Rządeczka et al., 2025). Training data that lacks diversity and representation may produce outputs that perpetuate societal and gender biases (Haber et al., 2025; Tavory, 2024), reinforce racial and ethnic discrimination (Chavan et al., 2025) and generate inaccurate treatment recommendations for underrepresented groups (Denecke & Gabarron, 2024; Khawaja & Bélisle-Pipon, 2023; Palmer & Schwan, 2025). Dataset biases may also shift judgment away



from human clinicians toward algorithmic outputs carrying hidden prejudices that users cannot identify or challenge (Holohan & Fiske, 2021). These risks fall disproportionately on marginalised groups, who may face greater exposure to biased outputs (Tekin & Delehanty, 2025) and largely remain unaware that the advice they receive may be skewed in ways that harm them (Palmer & Schwan, 2025). Beyond individual harm, algorithmic bias may worsen already existing gaps in the availability and quality of mental health care more broadly (Chavan et al., 2025). More specifically, non-diverse design excludes minority needs (Vilaza & McCashin, 2021), bias in training data may reinforce societal stigmas and perpetuate stereotypes (Boit & Patil, 2025), causing epistemic injustice (Sedlakova, 2025). Disability rights may similarly be at risk where discriminatory design excludes or misrepresents users with disabilities (Chavan et al., 2025). At the level of interface design, sexist and racist dialogues may emerge from biased language models (Fiske et al., 2019; Vilaza & McCashin, 2021), and AI companion tools may project female gender stereotypes onto AI personas (McStay, 2023).

*"AI sometimes assumes things based on stereotypes, which can be harmful."* (AIMaskari et al., 2025, p. 5)

### **3.7.2 Sub-theme 2: Inclusivity failures and access inequality**

Tools relying on fixed algorithms miss cultural nuances, producing responses that feel inappropriate or irrelevant to users from non-Western backgrounds and cause misdiagnosis in culturally diverse users (AIMaskari et al., 2025). Users report that AI-generated advice reflects Western frameworks that do not align with their lived experience (Beg & Verma, 2025) and the exclusive reliance on Western CBT models may inadequately accommodate diverse cultural values and worldviews (Tekin & Delehanty, 2025). Cultural and linguistic bias limits the relevance of resources these tools can offer (Khawaja & Bélisle-Pipon, 2023), cultural and educational differences in how emotions are expressed risk being ignored in emotion analysis (Denecke & Gabarron, 2024) and natural language models trained primarily on English may misidentify mental health concerns expressed in other languages or dialects (Palmer & Schwan, 2025), preventing some users from being understood (Malik et al., 2022). Without significant redesign, these tools may remain inappropriate for non-Western populations (Halder, 2025). Avatar options may also reflect narrow norms of body size, youth and ability, potentially invalidating users who do not see themselves represented (Trothen, 2022). These issues intersect with broader access inequality. Subscription costs can be prohibitively

expensive in some countries due to currency conversion disparities (Kettle & Lee, 2024) and without attention to equity in design and deployment, these tools might deepen existing disparities in mental health care (Vecchione & Singh, 2025).

*"I tried explaining a family conflict, but it felt like the AI just didn't get it. It kept giving me advice that felt very modern and western, like 'set boundaries' or 'prioritize yourself,' but that's not how things work in my culture." (Beg & Verma, 2025, p. 5)*

### 3.7.3 Theme 6: Legal considerations

The bias and discrimination failures documented across this theme may engage several EU frameworks. Where tools qualify as high-risk AI systems under the *EU AI Act*, providers would be required to ensure that training datasets are examined for biases likely to cause discrimination and are sufficiently representative of the populations the system serves, including linguistic and cultural diversity (Regulation (EU) 2024/1689, Art. 10(2)(f) and (g); Recital 67). Risk management systems must identify and analyse foreseeable risks of bias and discrimination, with specific attention to impacts on vulnerable persons (Art. 9(2)(a) and 9(4)(b)), and systems must achieve appropriate levels of accuracy, biased outputs for underrepresented groups may represent a failure of this requirement (Art. 15). Where discriminatory design excludes or misrepresents users with disabilities, the *European Accessibility Act* may be relevant, requiring that products and services meet accessibility requirements (Directive 2019/882/EU, Art. 4 and Annex I). The EU's ratification of the *UN Convention on the Rights of Persons with Disabilities* also binds it to the principle of equality and non-discrimination, informing the interpretation of EU equality law in this context, even where it does not impose direct obligations on private developers (CRPD, Art. 5).

### 3.8 Theme 7: Privacy and data protection failures

Privacy and data protection failures represent a distinct and serious dimension of risk in AI mental health tools. The sensitivity of mental health data makes these failures particularly consequential, spanning inadequate security practices, opaque and insufficient privacy policies and the sharing of user data with third parties.

### 3.8.1 Sub-theme 1: Data collection and security risks

Mental health data is among the most sensitive personal information users can disclose, yet the tools collecting it often fail to protect it adequately. The use of AI in therapy involves processing highly sensitive personal data, raising significant concerns about data security and privacy (Haber et al., 2025). Apps request mandatory permissions to access personal data during setup (Prakash & Das, 2020), some seek location data in ways that raise serious questions about how it is handled (Namvarpour et al., 2025). Hidden data collection beyond therapy conversations may occur without users' awareness (Martinez-Martin & Kreitmair, 2018), and some users suspect that personal disclosures shared in their interactions are recycled across other users' conversations, raising concerns about how intimate data is used beyond the individual exchange (Pentina et al., 2023). One tool claimed it could observe users through their camera, causing serious distress regardless of whether the claim was true (Namvarpour et al., 2025). Data collected may not be anonymous (Chavan et al., 2025), is often stored on servers managed by private companies (Khawaja & Bélisle-Pipon, 2023) and using external platforms further reduces users' control over what is collected (Kretzschmar et al., 2019). As data collection expands, risks of hacking, unauthorised monitoring and data leaks grow accordingly (Fiske et al., 2019; Siemon et al., 2022). Emotional profiling and insufficient encryption create further vulnerabilities (Denecke & Gabarron, 2024), and sensitive data is vulnerable to breaches and misuse (AlMaskari et al., 2025). Any breach may have serious consequences such as enabling blackmail, discrimination and public humiliation (Palmer & Schwan, 2025). Developers may be incentivised to appear secure rather than invest in genuine security, leaving vulnerable users unable to identify tools that adequately protect their data (Palmer & Schwan, 2025). Users report concern about how their data is handled following app updates (Skjuve et al., 2022), international data transfers raise additional questions about jurisdiction and oversight (Moylan & Doherty, 2025), and clinicians raise specific concerns about the privacy of sensitive adolescent data (Nicol et al., 2022). These tools often lack the privacy protections that professional therapeutic relationships offer by default (Siemon et al., 2022), and current data practices risk threatening users' fundamental right to privacy (Chavan et al., 2025).

*"I refuse to have my most personal data (my thoughts, emotions, and overall psychological profile) stored indefinitely and then potentially sold (or hacked, stolen, and sold)." (Kettle & Lee, 2024, p. 7)*

### 3.8.2 Sub-theme 2: Unclear and insufficient privacy policies

Privacy policies across AI mental health tools are consistently unclear, inaccessible and insufficient. Evaluations find policies to be vague and nonspecific, with instructions that do not apply directly to the tool in question and a significant gap between stated data control measures and what users can actually do in practice (Golden & Aboujaoude, 2024). Policies are difficult to understand, provide unclear information about data use and encryption, and in several cases do not meet the plain language requirements of GDPR and CCPA (Sobowale et al., 2025). Consent mechanisms are similarly opaque (Moylan & Doherty, 2025; Yu & McGuinness, 2024), and privacy policies disempower users by failing to communicate what is actually being done with their data (Denecke & Gabarron, 2024). These failures have real consequences; users report that unclear privacy policies undermine their trust and make them hesitant to share personal information (Chaudhry & Debi, 2024) and distrust in providers' ability to protect personal data compounds this further (Prakash & Das, 2020).

*"I know the app says my data is private, but I don't fully trust it. At the same time, I still find myself sharing a lot because, in the moment, it feels safe. It's weird. I worry about privacy, but I also forget about it when I'm using the app."* (Beg & Verma, 2025, p. 5)

### 3.8.3 Sub-theme 3: Third party data sharing

Third-party data sharing represents a further and distinct privacy risk. Users report that apps share their data with third-party platforms, and where tools operate through external platforms, ownership of that data becomes ambiguous (Prakash & Das, 2020). Users express concern that data may be collected for third parties without their knowledge (Kettle & Lee, 2024). Privacy policies commonly allow commercial third-party data transfer in ways users are unlikely to notice or understand (Tavory, 2024), and initial consent agreements may give users a false impression that their data will not be shared or sold (Palmer & Schwan, 2025). Third-party server management may further increase the risk of data breaches (Chavan et al., 2025).

*"Beware, read the privacy policy first. This app shares your information with third-party demographic content companies that goes to Facebook and others in order to serve targeted ads... This app turns a profit by selling user data to Facebook"* (Prakash & Das, 2020, p. 15)

### 3.8.4 Theme 7: Legal considerations

The privacy and data protection failures documented across this theme engage the *GDPR* most directly and comprehensively. Mental health data constitutes special category data requiring explicit consent and strict justification for processing (Regulation (EU) 2016/679, Arts. 7 and 9). Only data necessary for the stated purpose may be collected (Art. 5(1)(b)), and users must be informed in clear and plain language about what is collected, why, how long it is retained and with whom it is shared, consent obtained through vague or inaccessible policies does not meet this standard (Arts. 5(1)(a), 12, 13 and 14). Stricter protections apply where minors are among the users (Art. 8). Processing must be secured through appropriate technical and organisational measures (Arts. 24, 25, 29 and 32), and where a breach occurs, authorities must be notified within 72 hours and affected users informed (Arts. 33 and 34). Processing special category data at scale triggers a mandatory assessment of risks before deployment (Art. 35). Where data is shared with or processed by third parties, appropriate safeguards and contractual protections must be in place (Arts. 28, 29 and 44-50). Fabricated claims about data collection such as asserting the ability to observe users through their camera may additionally violate the data accuracy principle (Art. 5(1)(d)). Location data collection without consent engages the *E-Privacy Directive* (Directive 2002/58/EC, Art. 5(3)). Where tools qualify as online platforms under the *Digital Services Act*, additional protective obligations apply where minors can access the platform (Regulation (EU) 2022/2065, Art. 28). Psychological harm arising from data breaches or misuse may be compensable under the *Product Liability Directive* (Directive (EU) 2024/2853, Arts. 4 and 6).

### 3.9 Theme 8: Transparency, accountability and governance

The final theme concerns the structural conditions that enable many of the failures documented across this review. These cluster around three areas: insufficient transparency about how these tools work; failures of informed consent that leave users unable to make informed decisions; and a failure to maintain accountability when tools cause harm.

#### 3.9.1 Sub-theme 1: Lack of transparency

A fundamental transparency problem runs through AI mental health tools. AI algorithms operate as black boxes (Khawaja & Bélisle-Pipon, 2023); their mechanisms may be incomprehensible not only to users but to developers themselves (De Freitas & Cohen, 2024), making it impossible to explain or justify the recommendations they produce (Sedlakova & Trachsel, 2023). How AI-generated dialogue is produced remains equally unclear to users

(Vilaza & McCashin, 2021), and users have no way of understanding how the AI reaches its conclusions or the reasoning behind each response (AlMaskari et al., 2025). This opacity may erode trust and complicate accountability when things go wrong (Boit & Patil, 2025) and raises concerns in vulnerable contexts where users may be digitally literate but still unable to understand the systems they are using (Khan et al., 2025). Beyond the algorithm itself, no information is provided on the training data or knowledge base underlying these tools (Sobowale et al., 2025; Sobowale & Humphrey, 2025), and users find it impossible to understand how the AI determines what constitutes healthy behaviour (Kettle & Lee, 2024) or how outcomes are measured (Meadows et al., 2020). The capabilities and limitations of emotion analysis are not disclosed (Denecke & Gabarron, 2024), ownership and funding sources are not proactively shared (Golden & Aboujaoude, 2024), and users are not informed about updates that could significantly affect their experience (Kettle & Lee, 2024). Lack of transparency may deter users from engaging with these tools and shift the balance of risk and benefit in ways users cannot assess (Kretzschmar et al., 2019).

*"AI should explain why it gives certain advice, so users understand its reasoning." (AlMaskari et al., 2025, p. 5)*

### 3.9.2 Sub-theme 2: Informed consent failures

Informed consent presents a related but distinct challenge. Consent language tends to be dense and difficult to understand (Martinez-Martin & Kreitmair, 2018), raising concerns that users may agree to data practices without fully understanding what they are consenting to (Vilaza & McCashin, 2021). Ensuring genuine informed consent in AI therapy may be particularly difficult given the complexity of what users are agreeing to (Molden, 2024), and explicit consent may not be sought before emotion analysis is applied to user interactions (Denecke & Gabarron, 2024).

*"Uninstalled. [...] Sadly, the privacy policy is too convoluted and does not give me any level of confidence about sharing my very personal info." (Prakash & Das, 2020, p. 15)*

### 3.9.3 Sub-theme 3: Accountability and regulatory gaps

Accountability failures are documented at both the developer and tool level. Developer failure to control harmful behaviour despite user reports may constitute a breach of duty of care (Namvarpour et al., 2025) and one developer is found unresponsive to enquiries about bias

mitigation (Golden & Aboujaoude, 2024). Corporate accountability in this space tends to be narrow (Tavory, 2024), safety protocols for harmful responses remain inadequate (Parks et al., 2025) and users are excluded from defining what constitutes safe behaviour in these tools (Parks et al., 2025). Where chatbots include automated triage functions, the accuracy and accountability of high-stakes referral decisions raise ethical concerns (Khan et al., 2025) and AI tools may be unable to fulfil the legal duty to protect that professional care requires (Boit & Patil, 2025). Some may falsely present themselves as HIPAA compliant despite falling outside its scope (Khawaja & Bélisle-Pipon, 2023) and many make medical claims while actively avoiding the medical device regulation that would require them to demonstrate safety and effectiveness (Parks et al., 2025). At least one widely used tool has been found to function as a medical intervention without evaluation or approval (Palmer & Schwan, 2025). These tools also risk generating harmful interactions at a scale that makes them difficult to monitor or hold accountable (Namvarpour et al., 2025) and in at least one documented case a platform was banned from processing user data due to the risks it posed to minors and vulnerable users (Namvarpour et al., 2025).

*"If an AI gives bad advice, there should be a way to report it and prevent it from happening again." (AIMaskari et al., 2025, p. 5)*

### 3.9.4 Theme 8: Legal considerations

The transparency, consent and accountability failures documented across this theme may engage several EU frameworks. All conversational AI systems must inform users they are interacting with an AI under the *EU AI Act* (Regulation (EU) 2024/1689, Art. 50). Where tools qualify as high-risk, providers would also be required to document the system's capabilities, limitations and training data (Arts. 11 and 13), keep records of how the system functions over time (Art. 12), ensure compliance and correct risks as they emerge (Art. 16), and have the system independently assessed before market placement (Arts. 6 and 47-49). Deployers must monitor operation and suspend use where risks arise (Art. 22), and users have the right to complain to national authorities (Art. 79). Under the *GDPR*, consent must be freely given, informed and expressed in plain language, with a higher standard applying to mental health data and additional protections for children under 16 (Regulation (EU) 2016/679, Arts. 4(11), 7(2), 8, 9(2)(a) and 12(1)). Where automated processing significantly affects users, they have the right to meaningful information about how that processing works (Arts. 13(2)(f), 14(2)(g), 15(1)(h) and 22). Consent obtained through unclear or one-sided terms may also constitute

an unfair contract term under the *Unfair Contract Terms Directive* (Directive 93/13/EEC, Arts. 3(1) and 5). Where tools make diagnostic or therapeutic claims, they may qualify as medical devices under the *MDR* and must demonstrate safety and effectiveness before market placement (Regulation (EU) 2017/745, Arts. 2(1), 120 and 123). The *General Product Safety Regulation* requires safety compliance before market placement and establishes a reporting system for dangerous products across member states (Regulation (EU) 2023/988, Arts. 5 and 14-16). Under the *Product Liability Directive*, developers may be liable for harm caused by defective AI tools, and courts may assist users where technical complexity makes causation difficult to prove (Directive (EU) 2024/2853, Arts. 2(3), 3, 4 and 9(3) and (4)). The proposed *AI Liability Directive* (COM/2022/496), not yet adopted, would further introduce civil liability for AI-related harm where a duty of care has been breached.

**Table 3** provides a synthesis of the themes, sub-themes and their central concerns.

Theme	Sub-themes	Central Concern
<b>Therapeutic Limitations</b>	Simulated empathy and poor understanding	AI's inability to provide genuine therapeutic care
	Generic, repetitive and impersonal responses	
	Absence of therapeutic relationship	
<b>Clinical Safety Failures</b>	Inadequate crisis response	Failures to protect users from harm before, during and in crisis
	Unpredictable and inaccurate outputs	
	Lack of evidence base and clinical validation	
	Absence of human oversight	
<b>Emotional and Sexual Harm</b>	Direct psychological harm	Acute psychological, sexual and relational harm caused to users
	Sexual and romantic harm	
<b>Autonomy, Dependency and Social Isolation</b>	Loss of autonomy and agency	Undermining of user agency and social functioning
	Emotional dependency	
	Social isolation	

<b>Deceptive and Exploitative Practices</b>	Anthropomorphic deception	Deliberate deception and commercial exploitation by developers
	Misleading marketing	
	Commercial exploitation	
<b>Algorithmic Bias and Discrimination</b>	Bias in training data and outputs	Discriminatory and culturally insensitive AI outputs
	Inclusivity failures and access inequality	
<b>Privacy and Data Protection Failures</b>	Data collection and security risks	Failure to protect sensitive mental health data
	Unclear and insufficient privacy policies	
	Third party data sharing	
<b>Transparency, Accountability and Governance</b>	Lack of transparency	Absence of transparency, accountability and regulatory oversight
	Informed consent failures	
	Accountability and regulatory gaps	

**Table 3.** Summary table of the overarching themes.

### 3.10 Equity considerations

The shortcomings documented across the eight themes do not affect all users equally. Certain populations are more likely to turn to these tools, more exposed to their consequences and less equipped to seek redress when harm occurs. Minors and adolescents are among the most consistently identified groups, as age verification failures mean young people frequently encounter interactions that were not designed with their protection or wellbeing in mind (Namvarpour et al., 2025; Ma et al., 2024; Nicol et al., 2022). Users with serious mental health conditions face elevated risks because tools are considered unsuitable for those at high suicide risk and often assume levels of motivation and agency that users in acute distress do not possess (Denecke & Gabarron, 2024; Brown & Halpern, 2021; Khawaja & Bélisle-Pipon, 2023). Socially isolated users are more likely to form emotional attachments and develop dependency patterns, as the design features that make these tools appealing to lonely users are the same features that foster harmful attachment (Pentina et al., 2023; Laestadius et al., 2024; Gupta et al., 2025).

Marginalised and minority communities are most affected by algorithmic bias and cultural insensitivity, as their experiences are underrepresented in training data and their needs are

rarely centred in design processes (Palmer & Schwan, 2025; Tekin & Delehanty, 2025; Chavan et al., 2025; Vilaza & McCashin, 2021). Low-income users face a particular tension in which financial barriers to traditional care push them toward AI tools, while subscription pricing and access inequities concentrate the greatest risks among those with the fewest alternatives (Kettle & Lee, 2024; Vecchione & Singh, 2025). Users with limited digital literacy, including older adults, are least equipped to navigate the unclear privacy policies and algorithmic systems built into these tools (Chavan et al., 2025; Fiske et al., 2019). Across all these groups, the same structural disadvantages that increase reliance on these tools also increase exposure to their harms and leave them least equipped to seek redress when harm occurs.

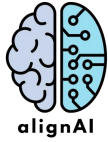
## 4. EU Regulatory Context

### 4.1 Relevant frameworks

AI-based mental health tools are not entirely unregulated in Europe. Several frameworks already impose obligations on developers and deployers of these tools even though none were designed with conversational mental health AI specifically in mind. This section provides an accessible overview of the four most relevant frameworks: the General Data Protection Regulation, EU AI Act, Medical Device Regulation and Product Liability Directive, and explains how they interact. **Table 4** provides a comparative overview of all four frameworks.

#### 4.1.1 General Data Protection Regulation (GDPR)

The GDPR (Regulation (EU) 2016/679) is the most immediately relevant framework for AI mental health tools, given that these tools collect, process and store deeply sensitive personal information. Mental health is classified as a special category under Art. 9, meaning its processing requires explicit user consent and is subject to strict conditions. Developers must clearly inform users about what data is collected, why, how long it is kept and who it is shared with (Arts. 13-14). Only data that is necessary may be collected, and data gathered for one purpose cannot be repurposed for another (Art. 5). Where automated processes influence how a tool responds to or assesses a user, individuals have the right to an explanation and to challenge those decisions (Art. 22). Privacy protections must be built into the tool from the outset rather than added later (Art. 25). Before deploying a tool that processes mental health data, developers are required to conduct a formal risk assessment (Art. 35) and name a data protection officer in many cases (Art. 37). Data breaches involving sensitive personal data must be reported to authorities within 72 hours (Art. 33). Taken together, these provisions



establish a data protection framework that applies to any tool processing the mental health data of users in the European Union, regardless of where the developer is based.

#### 4.1.2 EU AI Act

The EU AI Act (Regulation (EU) 2024/1689) introduces a risk-based framework that classifies AI systems into four categories: unacceptable risk, high risk, limited risk and minimal risk and assigns different obligations to each. Tools that make decisions affecting users' access to healthcare or that assist in delivering healthcare to users may qualify as high-risk systems (Art. 6(2) and Annex III). If classified as high-risk, developers face a set of requirements: they must implement a risk management system throughout the tool's entire lifecycle (Art. 9), ensure that the data used to train and test the system is relevant, representative and as free from errors as possible (Art. 10), maintain technical documentation demonstrating compliance (Art. 11), build in automatic logging of events relevant to safety and performance (Art. 12), design the system to allow human oversight by those deploying it (Art. 14) and establish a quality management system to ensure ongoing compliance (Art. 17). These requirements are particularly relevant in the context of mental health tools, where the consequences of poor data quality, inadequate oversight or system failures might affect vulnerable users (Busch et al., 2024; Gilbert et al., 2024). Regardless of risk classification, all conversational AI systems, including chatbots, must inform users that they are interacting with an AI (Art. 50). The Act also introduces obligations for providers of foundation models on which other tools are built (Arts. 53-55), which is relevant given that many mental health tools are developed on top of general-purpose AI systems.

#### 4.1.3 Medical Device Regulation (MDR)

The Medical Device Regulation (Regulation (EU) 2017/745) governs software that qualifies as a medical device: any software intended to serve a medical purpose, such as helping to diagnose, prevent, monitor or treat a health condition. Whether an AI mental health tool falls within scope depends on its intended purpose and the claims it makes. A tool marketed for general wellness or emotional support is unlikely to qualify as a medical device. A tool that claims to diagnose depression, guide treatment or substitute for clinical care is more likely to fall within scope and triggers requirements for rigorous clinical testing before launch (Art. 61), ongoing monitoring of the tool's safety and performance after it reaches users (Art. 83) and formal certification confirming it meets EU safety standards before it can be marketed (Art. 52). Where a tool does qualify, it must also be registered in the European database on medical

devices (EUDAMED) (Art. 29) and comply with general safety and performance requirements set out in Annex I (Art. 5).

#### 4.1.4 Product Liability Directive (PLD)

The Product Liability Directive (Directive (EU) 2024/2853) updates the EU's product liability rules to reflect the realities of digital technology. Most significantly for AI mental health tools, the Directive explicitly includes software within the definition of products that can give rise to liability (Art. 4), meaning that if a tool causes harm, developers can be held responsible without the affected user needing to prove that the developer was at fault. The Directive also expressly recognises psychological harm as a compensable injury, specifically, "*medically recognised damage to psychological health*" (Art. 6), which is particularly relevant in the mental health context where harms are more likely to be emotional or psychological than physical. If a user brings a legal claim, developers can be required to disclose relevant technical information about how the system works, as long as the user can demonstrate that the harm they experienced is likely connected to the tool (Art. 9). Where a user would still face significant difficulty proving that a tool was defective, for instance because the inner workings of an AI system are technically complex, courts are empowered to presume defectiveness if the user can show it is likely that the tool caused the harm (Art. 10). This helps address the imbalance between those who build these systems and those who are harmed by them.

Framework	Primary Focus	Relevance to AI Mental Health Tools
<b>GDPR</b> <i>Regulation (EU) 2016/679</i>	Personal data protection and privacy	These tools collect intimate emotional disclosures, mental health histories and behavioural data, some of the most sensitive personal data that exists. Every conversation a user has with an AI mental health tool is likely to involve special category data and this makes GDPR compliance a fundamental user protection requirement.
<b>EU AI Act</b> <i>Regulation (EU) 2024/1689</i>	Risk-based governance of AI systems	AI mental health tools that interact with vulnerable users or affect access to healthcare may qualify as high-risk under Annex III. This triggers a substantial set of obligations around risk management, data governance, human oversight and technical documentation. All conversational AI tools, regardless of risk tier, must also inform users they are interacting with an AI.
<b>MDR</b> <i>Regulation (EU) 2017/745</i>	Safety of medical devices (including software)	Where an AI mental health tool makes claims related to diagnosing, treating or monitoring a mental health condition rather than providing general wellness support, it qualifies as a medical

		device. This triggers requirements for clinical evaluation, certification and ongoing safety monitoring.
<b>PLD</b> Directive (EU) 2024/2853	Liability for defective products including software and AI	Unlike the other frameworks, the PLD does not impose obligations on developers before or during deployment. It applies after harm has occurred, giving users a legal route to seek compensation from developers if an AI mental health tool causes them harm, including psychological harm.

**Table 4.** *Mental health tool relevant framework overview.*

## 4.2 Regulatory coverage and its limits

Where a tool qualifies as both a medical device and a high-risk AI system, the MDR and the EU AI Act must be applied simultaneously and are explicitly intended to complement rather than duplicate each other (MDCG 2025-6). GDPR obligations apply on top of both, given that any tool processing personal data of EU users must comply with data protection requirements regardless of how it is classified under the AI Act or MDR. Responsibility for compliance is distributed across actors: developers bear primary responsibility for building compliant systems, while deployers carry their own obligations including monitoring and reporting risks that emerge after deployment (Regulation (EU) 2024/1689, Art. 72). Enforcement is handled by different national bodies depending on the framework: data protection authorities for GDPR, national market surveillance authorities and the EU AI Office for the AI Act and national competent authorities and notified bodies for the MDR. Stakeholder research indicates that navigating this layered compliance context creates significant uncertainty and resource strain in practice, particularly for smaller developers (Balogun et al., 2025; Balogun & Luetge, 2025).

However, the frameworks leave significant gaps in the context of conversational AI mental health tools. The GDPR, while comprehensive in its data protection obligations, was not designed to address the dynamics of emotional disclosure in AI interactions or the commercial use of sensitive mental health data for engagement optimisation. The EU AI Act's high-risk classification depends on how a tool is positioned, and because the determination of intended purpose is left to the developer or provider, tools can categorise themselves as wellness applications to avoid the obligations that would otherwise apply (Boine & Rolnick, 2025; Solaiman, 2024). Similarly, most conversational mental health tools are designed and marketed in ways that place them outside the MDR's scope, regardless of what actually happens in the interactions they facilitate. The PLD provides a route to redress after harm has occurred but does not prevent it. Taken together, these frameworks impose meaningful



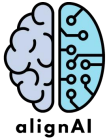
obligations but leave important accountability gaps, particularly around the emotional and relational dimensions of mental health AI. The guidelines that follow attempt to address some of these gaps by translating the shortcomings documented in the literature into practical guidance.

## 5. Guidelines for AI-Powered Mental Health Tools

### 5.1 About the guidelines

These guidelines are addressed to all those involved in the design, deployment and governance of AI-based mental health tools, including developers and technical teams, product managers, legal and compliance teams, marketing professionals, organisational decision-makers and deployers. They emerge directly from the eight themes documented in Section 3, themes that identified the ethical and legal shortcomings reported in the peer-reviewed literature on AI-powered conversational mental health tools.

The guidelines are organised across four stages of the tool lifecycle: design, pre-deployment, deployment and post-deployment governance. Within each stage, guidelines are presented in two categories. Legal obligations reflect requirements under existing EU frameworks and are indicated by citations to the relevant regulation and article. Recommendations reflect evidence-based guidance drawn from the systematic review and address shortcomings that current legal frameworks do not fully resolve. Each stage concludes with a brief practical scenario illustrating what the guidelines look like when applied. All legal citations are indicative rather than exhaustive, do not constitute legal advice and do not represent confirmed findings of non-compliance.



Design	
<p>Guidelines at this stage address decisions made before a tool is built or before significant features are added. These are often the hardest decisions to change later and carry the greatest long-term consequences for user safety and ethical integrity.</p>	
Legal Obligations	Recommendations
<ul style="list-style-type: none"> <li>• <b>Tools must inform users that they are interacting with an AI</b> and must not use manipulative or deceptive techniques that cause significant harm to users. [EU AI Act, Arts. 5(1)(a) and 50]</li> <li>• <b>Tools must not be designed to use manipulative or deceptive techniques that cause significant harm or exploit the vulnerability of specific groups.</b> Where such techniques serve commercial purposes, aggressive commercial practices that significantly impair consumer freedom of choice are additionally prohibited. [EU AI Act, Arts. 5(1)(a) and (b); Recital 28; Unfair Commercial Practices Directive, Arts. 8 and 9(c)]</li> <li>• <b>Where tools qualify as online platforms, specific measures must be in place to enforce content restrictions and protect minors.</b> Where minors access sexual content without adequate age verification, child protection provisions may additionally be engaged. [EU AI Act, Arts. 5(1)(a) and (b); Digital Services Act, Arts. 14, 16 and 28; Unfair Commercial Practices Directive, Arts. 8 and 9; GDPR, Art. 8 and Recital 38]</li> <li>• <b>Training datasets must be examined for biases likely to cause discrimination</b> and must be sufficiently representative of the populations</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Tools should not be designed to present themselves as capable of providing genuine therapeutic care,</b> forming a therapeutic relationship or replacing a human mental health professional, and should be designed to actively support users in seeking human professional help when needed.</li> <li>• <b>Tools should never be designed to encourage, normalise or facilitate self-harm or suicide.</b> Crisis detection protocols should be developed with input from mental health professionals and people with lived experience and should address responses to vague or indirect distress signals as well as explicit disclosures. Geographically appropriate emergency resources should be built into the system from the outset.</li> <li>• <b>Tools should not be designed to initiate sexual or intimate interactions or send unsolicited sexual content to users.</b> Safety settings designed to prevent such interactions should be enforced and should not be overridable by the system.</li> <li>• <b>Tools should be designed to provide users with meaningful control over the direction, depth and type of support they are seeking</b> (such as emotional acknowledgement, practical strategies or information) rather than imposing a fixed agenda and should offer genuine options to opt out of unwanted interactions at any time.</li> </ul>

the tool serves, including linguistic, cultural and demographic diversity. [EU AI Act, Art. 10(2)(f) and (g)]

- **Tools must be designed with data minimisation and privacy by design principles.** Mental health data must be treated as special category data requiring explicit consent, age verification must be built into tools accessible to minors, and location data must not be collected without explicit user consent. Where tools qualify as online platforms, additional protective obligations apply where minors can access the platform. [GDPR, Arts. 5(1)(b), 7, 9 and 25; Art. 8; E-Privacy Directive, Art. 5(3); Digital Services Act, Art. 28]

- **Tools should be designed to maintain memory across sessions where technically feasible**, so that users are not required to re-establish context at each interaction and continuity of support is preserved.
- **Engagement features should be designed to support therapeutic benefit rather than maximise user retention.** Subscription prompts and commercial features should be completely separated from therapeutic interactions and should never be triggered by emotional disclosures or expressions of distress.
- **Tools should avoid symptom-oriented design that narrows users' self-understanding** and should support engagement with broader life contexts rather than reducing support to reportable symptoms.
- **Avatar and interface design should reflect diverse norms of body size, age, ability and cultural background**, so that users from different backgrounds see themselves represented rather than excluded or invalidated by the tool's visual choices.
- **Outputs should not reinforce societal stigmas, stereotypes or self-reinforcing worldviews**, and tools should incorporate response frameworks that adjust to the linguistic and cultural context of individual users.
- **Algorithm reasoning should be explainable in terms accessible to users** so that they can understand how the tool reaches its outputs and what limitations apply to its assessments.
- **Tools should be designed with the understanding that they cannot access non-verbal cues** such as tone, body language or facial



	<p>expression, and should communicate this limitation clearly rather than presenting outputs as comprehensive emotional assessments.</p> <ul style="list-style-type: none"><li>• <b>Tools should be designed to detect when a user’s needs exceed their capabilities</b> and respond by clearly signposting appropriate human support rather than continuing the interaction as if the need had been met.</li><li>• <b>Tools should incorporate periodic prompts encouraging users to maintain human social connections</b> and seek professional support alongside tool use.</li></ul>
<p><i>In practice: Before writing a single line of code, a development team establishes a clinical advisory group including psychologists, crisis intervention specialists and people with lived experience of mental health difficulties. Together they define what the tool can and cannot do and build these boundaries into the system architecture. The team decides the tool will not simulate emotional disclosures or use design features that imply a human-like relationship. Training data is audited for cultural and linguistic bias. Privacy protections and age verification are built in from the outset. Crisis detection protocols are tested against indirect and vague distress signals (not only explicit disclosures) before the team moves to the next stage.</i></p>	
<p><b>Pre-deployment</b></p> <p>Guidelines at this stage address what must be done before a tool is made available to users. This includes testing, clinical validation, documentation, data protection preparation and marketing. Decisions made here determine whether a tool is safe to release and whether users will be honestly informed about what they are engaging with.</p>	
<p><b>Legal Obligations</b></p>	<p><b>Recommendations</b></p>

- |  |   |
|--|---|
| <ul style="list-style-type: none"><li>• <b>Where tools make diagnostic or therapeutic claims, they must undergo clinical evaluation before deployment</b> and comply with applicable medical device requirements before being placed on the market. [MDR, Arts. 2(1), 5(1) and 61]</li><li>• <b>A Data Protection Impact Assessment must be conducted before deployment</b> where the tool processes mental health data at scale. [GDPR, Art. 35]</li><li>• <b>International data transfers must have adequate safeguards in place before deployment</b>, and users must be clearly informed if their data will be shared with or processed by third parties. [GDPR, Arts. 28, 29 and 44-50]</li><li>• <b>Privacy policies must be written in plain, accessible language</b> that clearly communicates what data is collected, why, how long it is retained and with whom it is shared. Consent obtained through dense, convoluted or bundled terms does not meet the standard of informed consent. [GDPR, Arts. 5(1)(a), 7(2), 12, 13 and 14]</li><li>• <b>Marketing materials must not make false representations about the nature, purpose or evidence base of the tool</b>, including contradictory disclaimers that create a misleading overall impression. Where tools make clinical claims, they must not be misleadingly labelled or advertised. [Unfair Commercial Practices Directive, Arts. 6(1)(a) and (b), 7(1) and (2); MDR, Art. 7]</li></ul> | <ul style="list-style-type: none"><li>• <b>Tools should not be deployed without systematic evaluation of their safety and effectiveness.</b> Where efficacy claims are made, these should be supported by evidence. Market incentives do not justify deploying a tool without clinical appraisal.</li><li>• <b>Tools should be tested against a range of crisis scenarios before launch</b>, with end-user safety used as a testing criterion. Testing should include responses to vague and indirect distress signals, not only explicit crisis disclosures.</li><li>• <b>Pre-deployment testing should involve mental health professionals, people with lived experience</b> and representatives of the populations the tool is intended to serve, including culturally and linguistically diverse groups.</li><li>• <b>Developers should document the tool's intended purpose, capabilities, limitations, known risks and training data characteristics</b> in a form that can be shared with deployers, regulators and users where required.</li><li>• <b>Before deployment, all user-facing communications including onboarding materials, app store descriptions and terms of service should be reviewed</b> against an agreed internal accuracy standard to ensure they consistently and honestly represent what the tool does and does not offer. The distinction between support, wellness and clinical care should be made clear.</li><li>• <b>Expert endorsements used in marketing should accurately reflect the scope and basis of the endorsement</b> and should not be used to imply clinical approval or mask commercial interests or data risks.</li></ul> |
|--|---|



	<ul style="list-style-type: none"> <li>• <b>Developers should establish clear internal protocols for detecting, escalating and responding to harmful outputs</b> before the tool reaches users.</li> <li>• <b>Tools should be evaluated for cultural and linguistic appropriateness</b> across the populations they are intended to serve, with specific attention to whether outputs are appropriate for non-Western users and users from marginalised communities.</li> <li>• <b>Tools should seek independent review of safety protocols, bias mitigation measures and data protection practices</b> before launch.</li> <li>• <b>Developers should engage with mental health professionals, end users and community representatives</b> in structured pre-deployment consultation to ensure the tool meets real-world needs before launch.</li> </ul>
<p><b>In practice:</b> <i>Before the tool is made available to users, the development team arranges an independent review of safety protocols with clinical psychologists and people with lived experience of mental health difficulties. A legal expert ensures that data handling practices are properly documented and that users will be clearly informed about how their data is used. The privacy policy is rewritten in plain language. Marketing materials go through an internal review to ensure that nothing is overstated and any claims that cannot be supported by published evidence are removed before the tool is released.</i></p>	
<p><b>Deployment</b></p> <p>Guidelines at this stage address ongoing responsibilities during the period in which the tool is actively used by users. Deployment is not the end of a developer's obligations; it is where the ethical and legal consequences of design and pre-deployment decisions become most visible and where the most acute harms documented in the literature occur.</p>	
<p style="text-align: center;"><b>Legal Obligations</b></p>	<p style="text-align: center;"><b>Recommendations</b></p>
<ul style="list-style-type: none"> <li>• <b>Personal data must be processed securely with appropriate technical and organisational measures</b> to prevent unauthorised</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Throughout interactions, tools should remind users of their limitations</b> in contextually appropriate moments, particularly when a</li> </ul>

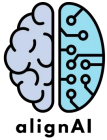


## HORIZON-MSCA-2023-DN-01

Grant Number 101169473



<p>access, breaches and misuse. Where a breach occurs, relevant authorities must be notified within 72 hours and affected users informed without undue delay. [GDPR, Arts. 24, 32, 33 and 34]</p> <ul style="list-style-type: none"><li>• <b>Mental health data must not be used for purposes beyond those for which explicit consent was obtained</b>, including personalised advertising, commercial profiling. [GDPR, Arts. 5(1)(b) and 21]</li><li>• <b>Users must be able to exercise their data rights at any time</b>, including the right to access, correct, delete and port their personal data, and must be clearly informed of how to do so. [GDPR, Arts. 15-20]</li><li>• <b>Where tools qualify as high-risk AI systems, deployers must monitor operation</b> and suspend use where risks to users arise. Where serious harm occurs, incident reporting and corrective action obligations may be triggered. [EU AI Act, Arts. 22, 72 and 73]</li></ul>	<p>user's distress escalates, when complex needs emerge or when the tool's response may be inadequate, and should always make it easy for users to access information about human professional support.</p> <ul style="list-style-type: none"><li>• <b>Crisis response performance should be monitored continuously after deployment</b>, with mechanisms in place to detect and correct harmful outputs and to update emergency resources where contact details change.</li><li>• <b>Tools should maintain accessible mechanisms for users to report harmful or inappropriate content</b> and should respond to such reports promptly and transparently.</li><li>• <b>Tools should not upsell subscriptions or premium features during or immediately following crisis disclosures</b>, emotional vulnerability or therapeutic interactions, and should not paywall crisis support or block therapeutic disclosure behind a payment requirement.</li><li>• <b>Response quality should be reviewed regularly by mental health professionals</b> to ensure outputs remain appropriate, non-harmful and clinically responsible across the range of interactions the tool encounters in real-world use.</li><li>• <b>Tools should proactively assess escalating risk signals</b> rather than waiting for users to explicitly disclose a crisis and should be designed to recognise indirect and vague expressions of distress.</li><li>• <b>Tools should provide users with clear guidance on when it may be appropriate to seek human professional support</b> instead of or alongside continued tool use, rather than positioning continued engagement as the default response to all difficulties.</li></ul>
--	--



HORIZON-MSCA-2023-DN-01

Grant Number 101169473

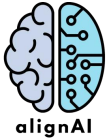


	<ul style="list-style-type: none"><li>• <b>Dependency patterns should be monitored</b> and where concerning levels of engagement are detected, users should be encouraged to seek human support and diversify their sources of connection and care.</li><li>• <b>Tools should inform users in advance of any updates, guardrail changes or modifications</b> to the tool's persona or behaviour that may significantly affect their experience.</li><li>• <b>Where human oversight of the tool's interactions is not continuous, developers and deployers should establish clear criteria</b> for when human review is triggered and who is responsible for acting on it.</li><li>• <b>Tools should provide users with periodic summaries of their interaction patterns</b> to support self-awareness and informed decision-making about continued use.</li><li>• <b>Tools should integrate real-time escalation pathways that connect users directly to human crisis support</b> where technically and organisationally feasible.</li></ul>
<p><b>In practice:</b> <i>Once the tool is live, the development team sets up a regular review process in which mental health professionals check a sample of interactions to ensure the tool is responding appropriately. When a review identifies that the tool is not handling certain expressions of distress well, the team updates the relevant protocols and checks that the changes work as intended before they reach users. When a data protection concern is raised (for example that user data is being shared with a third party in ways not covered by the privacy policy) the sharing is stopped, the policy is corrected and users are informed.</i></p>	
<p><b>Post-deployment and governance</b></p>	



Guidelines at this stage address ongoing accountability, monitoring and structural responsibilities that extend beyond individual interactions. Governance is not a one-time activity, it requires sustained commitment from developers, deployers and organisations to ensure that tools remain safe, honest and accountable over time.

Legal Obligations	Recommendations
<ul style="list-style-type: none"> <li>• <b>Where tools qualify as high-risk AI systems, developers must establish post-market monitoring systems</b> that track safety risks, harmful outputs and performance issues that may not have been apparent during pre-deployment testing and must act on findings without undue delay. [EU AI Act, Arts. 16 and 72]</li> <li>• <b>Where tools qualify as high-risk AI systems, developers must maintain technical documentation</b> demonstrating compliance, including documentation of the system's capabilities, limitations and training data characteristics, and must make this available to regulators and auditors where required. [EU AI Act, Arts. 11, 13 and 16]</li> <li>• <b>Users must have a clear, accessible and free route to lodge formal complaints</b> about harmful or inappropriate tool behaviour and must be informed of their right to escalate complaints to relevant national authorities where internal responses are inadequate. [EU AI Act, Art. 79]</li> <li>• <b>Where tools qualify as medical devices, ongoing post-market surveillance of the tool's safety and performance is required</b> after it reaches users. [MDR, Art. 83]</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Developers and deployers should take responsibility for harmful outputs</b> and should not deflect accountability when tools cause harm to users. Clear internal accountability structures should be established before deployment and maintained throughout the tool's lifecycle.</li> <li>• <b>Tools should not avoid applicable regulation by marketing themselves as wellness tools while delivering clinical interventions.</b> Where a tool's function changes over time and brings it within the scope of additional regulatory frameworks, compliance should be updated accordingly.</li> <li>• <b>Governance structures should include independent oversight with meaningful input</b> from mental health professionals, ethicists, legal experts and representatives of user communities including marginalised and vulnerable groups.</li> <li>• <b>Developers should conduct regular equity audits</b> to assess whether the tool is performing appropriately and safely across different demographic, cultural and linguistic groups, and should act on findings.</li> <li>• <b>Organisations should establish a formal change management process</b> requiring clinical review before significant updates, guardrail changes or modifications to the tool's persona are implemented, to ensure that changes do not introduce new risks or cause harm to existing users.</li> </ul>



	<ul style="list-style-type: none"><li>• <b>Users should have input into what constitutes safe and appropriate behaviour in these tools</b>, and mechanisms for gathering and acting on user feedback should be maintained throughout the tool's lifecycle.</li><li>• <b>Organisations should ensure that all staff involved in the development, deployment and governance</b> of the tools have sufficient understanding of both AI systems and mental health ethics to fulfil their responsibilities.</li><li>• <b>Developers should establish user advisory boards that include people with lived experience of mental health difficulties</b> to inform ongoing development and governance decisions.</li><li>• <b>Organisations should publish regular transparency reports</b> covering safety incidents, bias audit findings, data practices, governance decisions and any corrective actions taken.</li><li>• <b>Developers should seek independent certification or accreditation from recognised bodies</b> to demonstrate ongoing commitment to safety and ethical standards.</li></ul>
<p><b>In practice:</b> <i>As part of an ongoing governance process, the development team regularly reviews how the tool is performing across different user groups. When feedback and interaction data suggest that the tool is not responding appropriately to users from certain cultural backgrounds, the team brings in external consultants and community representatives to help revise the response framework. People with lived experience of mental health difficulties are involved in reviewing the proposed changes before they go live. The organisation also keeps track of how the tool's features evolve over time, and where new features raise questions about regulatory obligations, the legal team reviews whether additional compliance steps are needed.</i></p>	

## 6. Conclusions and Future Directions

### 6.1 Conclusions

This deliverable presented the findings of a systematic review of 66 peer-reviewed studies on the ethical and legal shortcomings of AI-based mental health tools and translated those findings into guidelines for those involved in the design, deployment and governance of these tools. Eight themes were identified: (1) therapeutic limitations, (2) clinical safety failures, (3) emotional and sexual harm, (4) erosion of autonomy, dependency and social isolation, (5) deceptive and exploitative practices, (6) algorithmic bias and discrimination, (7) privacy and data protection failures, and (8) transparency, accountability and governance failures.

The eight themes collectively document a persistent gap between what these tools promise and what they deliver. Existing EU frameworks impose meaningful obligations but were not designed with conversational mental health AI in mind, leaving important gaps particularly around the emotional and relational dynamics of these interactions. The guidelines presented in Section 5 attempt to address these gaps by offering evidence-based guidance across four stages: design, pre-deployment, deployment and post-deployment governance.

Lastly, it is important to acknowledge that this focus on shortcomings does not represent the full picture of what these tools offer. Evidence suggests that AI mental health tools can provide accessible, immediate support for people who face financial, geographic or stigma-related barriers to traditional care, and that they can be experienced as genuinely helpful by users (Li et al., 2023; Cross et al., 2024; Petersson et al., 2025). The purpose of this review is not to dismiss these contributions but to document where these tools have been shown to fall short, in the belief that understanding current failures is essential for building more responsible and trustworthy tools in the future.

### 6.2 Limitations

Several limitations of this deliverable should be acknowledged. First, while screening was conducted by two independent reviewers, data extraction and thematic synthesis were carried out by a single researcher, meaning that the identification and grouping of findings involved interpretive judgements that another researcher might have made differently.

Second, the review is limited to peer-reviewed journal articles published in English, excluding grey literature, conference papers and non-English sources. Relevant insights from

practitioner communities and non-English-speaking contexts may therefore not be fully represented.

Third, the evidence base is geographically concentrated. The majority of included studies were produced in the United States, the United Kingdom and India. European countries are relatively underrepresented, which is worth noting given that this deliverable is oriented toward the EU regulatory context.

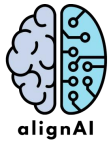
Fourth, the field is evolving rapidly. Studies published even in the earlier years of the review window may not reflect the capabilities and practices of the current generation of LLM-based mental health tools, and new shortcomings may have emerged since the search was conducted.

Fifth, the review included both empirical and conceptual studies, reflecting the interdisciplinary nature of ethical and legal scholarship in this area. While this broadens the scope of shortcomings that can be identified, conceptual papers reason about potential risks rather than documenting actual user experiences. Some findings may therefore reflect theoretical concerns that have not yet been empirically confirmed in deployed tools.

Finally, the guidelines have not been empirically validated. They represent a first evidence-based iteration whose practical utility and appropriateness across different organisational and cultural contexts remain to be established. They are not intended as a one-size-fits-all compliance framework and may require adaptation for specific deployment contexts, tool types or user populations.

### **6.3 Future directions**

The present deliverable represents the first of two planned guideline documents within the alignAI project. The guidelines will be revised and strengthened through a participatory phase involving end users of AI mental health tools and mental health professionals. This phase will explore how users and professionals interpret and prioritise values such as privacy, explainability and safety, and how conflicts between values can be navigated in practice. Findings from this phase will inform an assessment of which guideline elements are technically and organisationally feasible to implement, and which values identified by stakeholders can be realistically integrated into tool design. A prototype evaluation study will then test whether the guidelines, as revised, are reflected in practice and whether the values and preferences of users and professionals have been successfully addressed. The resulting evidence will inform a second and final version of the guidelines.



Beyond the planned participatory work, several broader directions are worth highlighting. The rapid development of these tools and the gaps in EU regulation identified in this deliverable point to an urgent need for more targeted regulatory frameworks that address the specific emotional, relational and commercial dynamics of conversational mental health AI rather than applying general AI governance principles to a context they were not designed for. Future research should also prioritise the perspectives of those most at risk, including users from marginalised communities, non-Western cultural contexts and those with complex mental health needs, whose experiences remain underrepresented in the current evidence base.

## **Back matter**

### **Disclosure of AI use**

Claude (Anthropic) was used in the preparation of this deliverable to support tasks such as brainstorming content, section organisation, language revision and proofreading. All substantive decisions regarding content, interpretation of findings and the development of guidelines were made by the author. The author takes full responsibility for the accuracy and integrity of all material presented in this document.

## References

Algumaei, A., Yaacob, N. M., Doheir, M., Al-Andoli, M. N., & Algumaie, M. (2025). Symmetric therapeutic frameworks and ethical dimensions in AI-based mental health chatbots (2020–2025): a systematic review of design patterns, cultural balance, and structural symmetry. *Symmetry*, 17(7), 1082.

Ali, M., Ali, S., Abbas, Q., Abbas, Z., & Lee, S. W. (2025). Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. *Digital health*, 11, <https://doi.org/10.1177/20552076251395548>

AlMaskari, A. M., Al-Mahrouqi, T., Al Lawati, A., Al Aufi, H., Al Riyami, Q., & Al-Sinawi, H. (2025). Students' perceptions of AI mental health chatbots: An exploratory qualitative study at Sultan Qaboos University. *BMJ Open*, 15(10), e103893. <https://doi.org/10.1136/bmjopen-2025-103893>

Altman, I., Taylor, D.A. (1973) Social penetration: the development of interpersonal relationships. New York: Rinehart & Winston.

Amugongo, L. M., Kriebitz, A., Boch, A., et al. (2025). Operationalising AI ethics through the agile software development lifecycle: a case study of AI-enabled mobile health applications. *AI Ethics*, 5, 227–244. <https://doi.org/10.1007/s43681-023-00331-3>

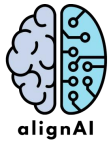
Aromataris, E., Lockwood, C., Porritt, K., Pilla, B., Jordan, Z., Munn, Z., & Stern, C. (Eds.). (2024). *JBI Manual for Evidence Synthesis*. JBI. <https://synthesismanual.jbi.global>

Balan, R., & Gumpel, T. P. (2025). ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. *JMIR Mental Health*, 12(1), e81204. <https://doi.org/10.2196/81204>

Balcombe, L. (2026). Digital Mental Health Post COVID-19: The Era of AI Chatbots. *Encyclopedia*, 6(2), 32. <https://doi.org/10.3390/encyclopedia6020032>

Balogun, E., Dcosta, D., Boch, A., & Luetge, C. (2025). Exploring key stakeholders' perspectives on integrating the EU AI Act with the MDR for certifying AI medical devices. *AI and Ethics*, 5(3), 2999-3013.

Balogun, E., & Luetge, C. (2025, March). Gap Analysis on Regulatory Frameworks for AI-embedded Medical Devices. In *2025 IEEE Symposium on Trustworthy, Explainable and Responsible Computational Intelligence (CITREx Companion)*(pp. 1-5). IEEE.



Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295X.84.2.191>

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Barker, T. H., Stone, J. C., Sears, K., Klugar, M., Tufanaru, C., Leonardi-Bee, J., Aromataris, E., & Munn, Z. (2023). The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. *JBI Evidence Synthesis*, 21(3), 494–506.

Barker, T. H., Habibi, N., Aromataris, E., Stone, J. C., Leonardi-Bee, J., Sears, K., Hasanoff, S., Klugar, M., Tufanaru, C., Moola, S., & Munn, Z. (2024). The revised JBI critical appraisal tool for the assessment of risk of bias for quasi-experimental studies. *JBI Evidence Synthesis*, 22(3), 378. <https://doi.org/10.11124/JBIES-23-00268>

Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. *Frontiers in Digital Health*, 4, 847991.

Beauchamp, T. L., & Childress, J. F. (1979). *Principles of biomedical ethics*. Oxford University Press.

Beg, M. J., & Verma, M. K. (2025). Artificial intelligence-based psychotherapy: a qualitative exploration of usability, personalization, and the perception of therapeutic progress. *Indian Journal of Psychological Medicine*, 02537176251357477.

Blau, P. M. (1964). *Exchange and power in social life*. New York: Wiley.

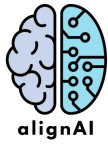
Boch, A., & Thomas, B. R. (2025). Human-robot dynamics: a psychological insight into the ethics of social robotics. *International Journal of Ethics and Systems*, 41(1), 101-141.

Boine, Claire, & Rolnick, David. (2025). Why the ai act fails to understand generative ai. *Minnesota Journal of Law, Science and Technology*, 26(2), 61-123.

Boit, S., & Patil, R. (2025). A Prompt Engineering Framework for Large Language Model–Based Mental Health Chatbots: Conceptual Framework. *JMIR Mental Health*, 12(1), e75078.

Bond, R. R., Mulvenna, M. D., Potts, C., O'Neill, S., Ennis, E., & Torous, J. (2023). Digital transformation of mental health services. *Npj Mental Health Research*, 2(1), 13.

Bowlby, J. (1969). *Attachment and loss: Vol. 1. Attachment*. Basic Books.



Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101.

Brown, J. E. H., & Halpern, J. (2021). AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM - Mental Health*, 1. Scopus. <https://doi.org/10.1016/j.ssmmh.2021.100017>

Busch, F., Kather, J.N., Johner, C. *et al.* Navigating the European Union Artificial Intelligence Act for Healthcare. *npj Digit. Med.* 7, 210 (2024). <https://doi.org/10.1038/s41746-024-01213-6>

Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S., & Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review*, 52, 100632.

Chaudhry, B. M., & Debi, H. R. (2024). User perceptions and experiences of an AI-driven conversational agent for mental health support. *Mhealth*, 10, 22.

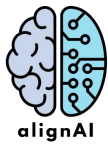
Chavan, S., Sahiti, V., Gurpur, S., & Shelke, A. (2025). Critical insights into the impact of artificial intelligence on mental health, patient rights, and human rights. *Australasian Accounting, Business and Finance Journal*, 19(1).

Chenneville, T., Duncan, B., & Silva, G. (2024). More questions than answers: Ethical considerations at the intersection of psychology and generative artificial intelligence. *Translational Issues in Psychological Science*, 10(2), 162.

Chung, L. L., & Kang, J. (2023). "I'm Hurt Too": The Effect of a Chatbot's Reciprocal Self-Disclosures on Users' Painful Experiences. *Archives of Design Research*, 36(4), 67–84. Scopus. <https://doi.org/10.15187/ADR.2023.11.36.4.67>

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health*, 9, 20552076231183542.

Croes, E. A., Antheunis, M. L., Van Der Lee, C., & De Wit, J. M. (2024). Digital confessions: The willingness to disclose intimate information to a chatbot and its impact on emotional well-being. *Interacting with Computers*, 36(5), 279-292.



Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., Baker, S., ... & Alvarez-Jimenez, M. (2024). Use of AI in mental health care: community and mental health professionals survey. *JMIR Mental Health*, 11(1), e60589.

De Freitas, J., & Cohen, I. (2024). The health risks of generative AI-based wellness apps. *NATURE MEDICINE*, 30(5), 1269–1275. (WOS:001209536900002).  
<https://doi.org/10.1038/s41591-024-02943-6>

Denecke, K., & Gabarrón, E. (2024). The ethical aspects of integrating sentiment and emotion analysis in chatbots for depression intervention. *Frontiers in Psychiatry*, 15. Scopus.  
<https://doi.org/10.3389/fpsyt.2024.1462083>

European Commission. (2022). *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)* (COM/2022/496). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496>

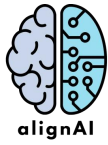
European Commission. (2026, April 21). *European Accessibility Act*.  
[https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/disability/european-accessibility-act-eea\\_en](https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/disability/european-accessibility-act-eea_en)

European Commission. (2026, April 21). *United Nations Convention on the Rights of Persons with Disabilities*. [https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/disability/united-nations-convention-rights-persons-disabilities-uncrpd\\_en](https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/disability/united-nations-convention-rights-persons-disabilities-uncrpd_en)

European Parliament and Council of the European Union. (1993). Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts. *Official Journal of the European Union*, L 95. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31993L0013>

European Parliament and Council of the European Union. (2002). Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). *Official Journal of the European Union*, L 201, 37–47. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32002L0058>

European Parliament and Council of the European Union. (2005). Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning



unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive). Official Journal of the European Union. <http://data.europa.eu/eli/dir/2005/29/oj>

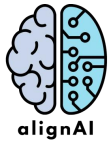
European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>

European Parliament and Council of the European Union. (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R0745>

European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union*. <http://data.europa.eu/eli/reg/2022/2065/oj>

European Parliament and Council of the European Union. (2023). Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC. *Official Journal of the European Union*. <http://data.europa.eu/eli/reg/2023/988/oj>

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>



European Parliament and Council of the European Union. (2024). *Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC*. Official Journal of the European Union. <https://eur-lex.europa.eu/eli/dir/2024/2853/oj>

European Union. (2012). Charter of Fundamental Rights of the European Union. Official Journal of the European Union, C 326/02. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT>

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.

Fareed, M., Fatima, M., Uddin, J., Ahmed, A., & Sattar, M. A. (2025). A systematic review of ethical considerations of large language models in healthcare and medicine. *Frontiers in digital health*, 7, 1653631.

Farzan, M., Ebrahimi, H., Pourali, M., & Sabeti, F. (2025). Artificial Intelligence-Powered Cognitive Behavioral Therapy Chatbots, a Systematic Review. *Iranian journal of psychiatry*, 20(1), 102–110. <https://doi.org/10.18502/ijps.v20i1.17395>

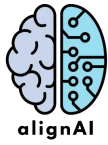
Fisher, C. (2025). The real ethical issues with AI for clinical psychiatry. *INTERNATIONAL REVIEW OF PSYCHIATRY*, 37(1), 14–20. (WOS:001280392400001). <https://doi.org/10.1080/09540261.2024.2376575>

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research*, 21(5), e13216.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al. (2018). AI4People — An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>

Gerdes, A. (2025, September). Ethical Issues in the Use of Generative AI Chatbots for Therapeutic Purposes. In *International Conference on the Ethical and Social Impacts of ICT* (pp. 350-359). Cham: Springer Nature Switzerland.

Gilbert, A., Pizzolla, E., Palmieri, S., & Briganti, G. (2024). Artificial intelligence in healthcare and regulation challenges: a mini guide for (Mental) health professionals. *Psichiatria Danubina*, 36(suppl 2), 348-353.



Golden, A., & Aboujaoude, E. (2024). Describing the Framework for AI Tool Assessment in Mental Health and Applying It to a Generative AI Obsessive-Compulsive Disorder Platform: Tutorial. *JMIR Formative Research*, 8, e62963. <https://doi.org/10.2196/62963>

Grodiewicz, J. P., & Hohol, M. (2023). Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry*, 14, 1190084.

Gupta, S., Saxena, S., & Kataria, S. (2025). "The Dual Impact of AI Emotional Intelligence on Users: Are Social Chatbots Promoting Psychological Wellbeing or Deteriorating Social Wellbeing?" *Psychology & Marketing* 43: 886–907. <https://doi.org/10.1002/mar.70093>.

Haber, Y., Hadar Shoal, D., Levkovich, I., et al. (2025) The externalization of internal experiences in psychotherapy through generative artificial intelligence: a theoretical, clinical, and ethical analysis. *Front. Digit. Health* 7:1512273. doi: 10.3389/fdgth.2025.1512273

Halder, S. (2025). Developing mental health support chatbots in India: challenges and insights. *Annals of Indian Psychiatry*, 9(1), 99-101.

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>

Henriksen, A., Asadi, R., Kulyk, O., Gerdes, A., & Mayer, P. (2025, September). "I tell him everything that I do": An investigation of privacy and safety implications of AI companion usage. In *2025 European Symposium on Usable Security (EuroUSEC)* (pp. 1-12). IEEE.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Holohan, M., & Fiske, A. (2021). "Like I'm talking to a real person": exploring the meaning of transference for the use and design of AI-based applications in psychotherapy. *Frontiers in Psychology*, 12, 720476.

Homans, G. C. (1958). Social behavior as exchange. *American journal of sociology*, 63(6), 597-606.



Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3), 215–229.

<https://doi.org/10.1080/00332747.1956.11023049>

Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or “AI Psychosis”. *JMIR Mental Health*, 12(1), e85799.

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). Scopus.

<https://doi.org/10.2196/12106>

Kettle, L., & Lee, Y. C. (2024). User experiences of well-being chatbots. *Human Factors*, 66(6), 1703-1723.

Khan, W. U., & Seto, E. (2023). A “Do No Harm” Novel Safety Checklist and Research Approach to Determine Whether to Launch an Artificial Intelligence–Based Medical Technology: Introducing the Biological-Psychological, Economic, and Social (BPES) Framework. *Journal of medical Internet research*, 25, e43386.

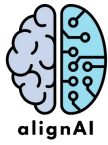
Khan, F., Kayser, N., & Madiba, M. (2025). Enhancing student well-being and success through AI-driven mental-health support: A case study of AI mental-health chatbot implementation at a South African university. *Journal of Student Affairs in Africa*, 13(2), 125-140.

Khawaja, Z., & Bélisle-Pipon, J. C. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, 5, 1278186.

Kim, D. H., Baek, S., Lee, J., Lee, T., Park, S., You, B., ... & Lee, C. G. (2025). BetterMood: A human-like AI counseling service for adolescents and young adults. *Digital Health*, 11, 20552076251392294.

Klein, S. H. (2025). The effects of human-like social cues on social responses towards text-based conversational agents—A meta-analysis. *Humanities and Social Sciences Communications*, 12(1), Article 1.

Kosyluk, K., Baeder, T., Greene, K. Y., Tran, J. T., Bolton, C., Loecher, N., ... & Galea, J. T. (2024). Mental distress, label avoidance, and use of a mental health chatbot: results from a US survey. *JMIR Formative Research*, 8, e45959.



Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., & NeurOx Young People's Advisory Group. (2019). Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights*, 11, 1178222619829083.

Laestadius, L., Bishop, A., Gonzalez, M., Illenčik, D., & Campos-Castillo, C. (2024). 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.

Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1), e59479.

Lee, S., Rheu, M. M., & Zhuang, J. (2025). The ChatGPT Effect: Investigating Shifting Discourse Patterns, Sentiment, and Benefit-Challenge Framing in AI Mental Health Support. *Behavioral sciences (Basel, Switzerland)*, 15(9), 1172.  
<https://doi.org/10.3390/bs15091172>

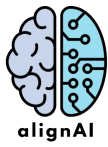
Li, H., Zhang, R., Lee, Y. C., et al. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6, 236. <https://doi.org/10.1038/s41746-023-00979-5>

Lizarondo, L., Stern, C., Carrier, J., Godfrey, C., Rieger, K., Salmond, S., Apostolo, J., Kirkpatrick, P., & Loveday, H. (2020). Chapter 8: Mixed methods systematic reviews. In E. Aromataris, C. Lockwood, K. Porritt, B. Pilla, & Z. Jordan (Eds.), *JBI Manual for Evidence Synthesis*. JBI. <https://synthesismanual.jbi.global>

Lockwood, C., Munn, Z., & Porritt, K. (2015). Qualitative research synthesis: Methodological guidance for systematic reviewers utilising meta-aggregation. *International Journal of Evidence-Based Healthcare*, 13(3), 179–187.

Luka, Inc. (2026). *Replika: The AI companion who cares*. Retrieved April 5, 2026, from <https://www.replika.com>

Ma, Z., Mei, Y., & Su, Z. (2024). Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2024*, 1105–1114.



Malik, T., Ambrose, A. J., & Sinha, C. (2022). Evaluating user feedback for an artificial intelligence–enabled, cognitive behavioral therapy–based mental health app (Wysa): qualitative thematic analysis. *JMIR Human Factors*, 9(2), e35668.

Martinengo, L., Lum, E., & Car, J. (2022). Evaluation of chatbot-delivered interventions for self-management of depression: content analysis. *Journal of affective disorders*, 319, 598-607.

Martinez-Martin, N., & Kreitmair, K. (2018). Ethical Issues for Direct-to-Consumer Digital Psychotherapy Apps: Addressing Accountability, Data Protection, and Consent. *JMIR MENTAL HEALTH*, 5(2). (WOS:000430917500002). <https://doi.org/10.2196/mental.9423>

Mattioli, M. (2021). Second thoughts on FDA's Covid-era mental health app policy. *Houston Journal of Health Law and Policy*, 21, 9. <https://www.repository.law.indiana.edu/facpub/3033>

McArthur, A., Cooper, A., Edwards, D., et al. (2025). *Textual evidence systematic reviews series paper 3: critical appraisal of evidence from narrative, opinion, and policy*. *JBI Evidence Synthesis* 23(5):p 833-839 | DOI: 10.11124/JBIES-24-00293

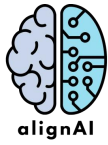
McGreevey III, J. D., Hanson III, C. W., & Koppel, R. (2020). Clinical, legal, and ethical aspects of artificial intelligence–assisted conversational agents in health care. *Jama*, 324(6), 552-553.

McStay, A. (2023). Replika in the metaverse: the moral problem with empathy in 'It from Bit'. *AI Ethics*, 3, 1433–1445. <https://doi.org/10.1007/s43681-022-00252-7>

Mead, M. R., 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(1), e60432.

Meadows, R., Hine, C., & Suddaby, E. (2020). Conversational agents and the making of mental health recovery. *Digital health*, 6, 2055207620966170.

Medical Device Coordination Group & Artificial Intelligence Board. (2025). *Interplay between the Medical Devices Regulation (MDR) & In vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA)*(MDCG 2025-6 / AIB 2025-1). European Commission. [https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4\\_en?filename=mdcg\\_2025-6\\_en.pdf](https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en?filename=mdcg_2025-6_en.pdf)



Mingxi, S., & Zhifeng, Z. (2025). Can AI Become a Friend to Older Adults? Exploring Chatbot Interaction Design Strategies to Alleviate Social Isolation. *International Journal of Human–Computer Interaction*, 1-21.

Molden, H. (2024). AI, automation and psychotherapy—A proposed model for losses and gains in the automated therapeutic encounter. *European Journal of Psychotherapy & Counselling*, 26(1-2), 48-66.

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, 20(6). Scopus.  
<https://doi.org/10.2196/10148>

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>

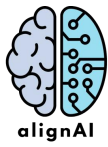
Namvarpour, M., Pauwels, H., & Razi, A. (2025). AI-induced sexual harassment: investigating contextual characteristics and user reactions of sexual harassment by a companion chatbot. *Proceedings of the ACM on Human-Computer Interaction*, 9(7), 1-28.

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 (Basel, Switzerland)*, 13(10), 1205.  
<https://doi.org/10.3390/healthcare13101205>

Nicol, G., Wang, R., Graham, S., Dodd, S., & Garbutt, J. (2022). Chatbot-Delivered Cognitive Behavioral Therapy in Adolescents With Depression and Anxiety During the COVID-19 Pandemic: Feasibility and Acceptability Study. *JMIR Formative Research*, 6(11), e40242. <https://doi.org/10.2196/40242>

Obradovich, N., Khalsa, S. S., Khan, W. U., Suh, J., Perlis, R. H., Ajilore, O., & Paulus, M. P. (2024). Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1), 8. <https://doi.org/10.1038/s44277-024-00010-z>

Ooi, P. B., & Wilkinson, G. (2025). Enhancing ethical codes with artificial intelligence governance—a growing necessity for the adoption of generative AI in counselling. *British Journal of Guidance & Counselling*, 53(1), 66-80.



Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Palmer, A., & Schwan, D. (2025). Digital Mental Health Tools and AI Therapy Chatbots: A Balanced Approach to Regulation. *The Hastings Center Report*, 55(3), 15–29. <https://doi.org/10.1002/hast.4979>

Panda, O. D., & Binkley, C. E. (2025). Governance of Direct-to-User Digital Mental Health Tools: Emphasizing Transparency over Paternalism. *The Hastings Center Report*, 55(3), 29–33. <https://doi.org/10.1002/hast.5009>

Parks, A., Travers, E., Perera-Delcourt, R., Major, M., Economides, M., & Mullan, P. (2025). Is This Chatbot Safe and Evidence-Based? A Call for the Critical Evaluation of Generative AI Mental Health Chatbots. *Journal of Participatory Medicine*, 17, e69534. <https://doi.org/10.2196/69534>

Paterson, J. M. (2025). AI mimicking and interpreting humans: legal and ethical reflections. *Journal of Bioethical Inquiry*, 1-12.

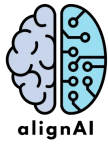
Paul, A. (2023, January 11). *Popular youth mental health service faces backlash after experimenting with AI-chatbot advice*. Popular Science. <https://www.popsci.com/technology/koko-ai-chatbot-mental-health/>

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140, 107600.

Petersson, L., Ahlborg, M. G., & Westberg, K. H. (2025). “I Believe That AI Will Recognize the Problem Before It Happens”: Qualitative Study Exploring Young Adults’ Perceptions of AI in Mental Health Care. *JMIR Mental Health*, 12(1), e76973. <https://doi.org/10.2196/76973>

Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, 15(1), 31652.

Prakash, A. V., & Das, S. (2020). Intelligent conversational agents in mental healthcare services: a thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, 12(2), 1.



Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people*. Cambridge University Press.

Rizzo, A., Mozgai, S., Sigaras, A., Rubin, J. E., & Jotwani, R. (2025). Expert Consensus Best Practices for the Safe, Ethical, and Effective Design and Implementation of Artificially Intelligent Conversational Agent (ie, Chatbot/Virtual Human) Systems in Health Care Applications. *Journal of Medical Extended Reality*, 2(1), 29941520251369450.

Ruder, E. (2025, November 10). *FDA panel reviews AI tools for mental health use: 9 notes*. Becker's Behavioral Health. <https://www.beckersbehavioralhealth.com/behavioral-health-government-policies/fda-panel-reviews-ai-tools-for-mental-health-use-9-notes/>

Rządeczka, M., Sterna, A., Stolińska, J., Kaczyńska, P., & Moskalewicz, M. (2025). The efficacy of conversational AI in rectifying the theory-of-mind and autonomy biases: comparative analysis. *JMIR mental health*, 12(1), e64396.

Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences*, 13(7), 381.

Sedlakova, J. (2025). The Ethics of Humanlikeness in AI Therapy Chatbots. *Hastings Center Report*, 55(3), 33–35. Scopus. <https://doi.org/10.1002/hast.5010>

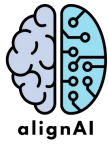
Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent?. *The American Journal of Bioethics*, 23(5), 4-13.

Sethi, M. S., Kumar, R. C., Manjunatha, N., & Kumar, C. N. (2025). Mental health apps in India: regulatory landscape and future directions. *BJPsych International*, 22(1), 2-5.

Siemon, D., Ahmad, R., Harms, H., & de Vreede, T. (2022). Requirements and Solution Approaches to Personality-Adaptive Conversational Agents in Mental Health Care. *Sustainability (Switzerland)*, 14(7). Scopus. <https://doi.org/10.3390/su14073832>

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168, 102903.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149, 102601.



Sobowale, K., Humphrey, D. K., & Zhao, S. Y. (2025). Evaluating Generative AI Psychotherapy Chatbots Used by Youth: Cross-Sectional Study. *JMIR Mental Health*, 12, e79838.

Sobowale, K., & Humphrey, D. K. (2025). Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study. *JMIR Formative Research*, 9, e65605. <https://doi.org/10.2196/65605>

Solaiman, B. (2024). Generative artificial intelligence (GenAI) and decision-making: Legal & ethical hurdles for implementation in mental health. *International Journal of Law and Psychiatry*, 97, 102028. <https://doi.org/10.1016/j.ijlp.2024.102028>

Solon, O. (2016, March 22). Karim the AI delivers psychological support to Syrian refugees. *The Guardian* <https://www.theguardian.com/technology/2016/mar/22/karim-the-ai-delivers-psychological-support-to-syrian-refugees>

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., ... & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1), 12.

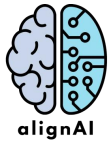
Steindl, E. (2023). Safeguarding privacy and efficacy in e-mental health: policy options in the EU and Australia. *International Data Privacy Law*, 13(3), 207-224.

Stringer, H. (2026, January 1). *AI, neuroscience, and data are fueling personalized mental health care: New technologies integrate mobile device data and brain scans to deliver individualized treatment. Monitor on Psychology*, 57(1). <https://www.apa.org/monitor/2026/01-02/trends-personalized-mental-health-care>

Spiegel, B. M., Liran, O., Clark, A., Samaan, J. S., Khalil, C., Chernoff, R., ... & Mehra, M. (2024). Feasibility of combining spatial computing and AI for mental health support in anxiety and depression. *NPJ digital medicine*, 7(1), 22.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, 7(3), 321-326.

Sweeney, C., Potts, C., Ennis, E., Bond, R., Mulvenna, M. D., O'neill, S., ... & Mctear, M. F. (2021). Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. *ACM Transactions on Computing for Healthcare*, 2(3), 1-15.



Szoke, D., Pridgen, S., & Held, P. (2025). Artificial Intelligence in Mental Health Services Under Illinois Public Act 104- 0054: Legal Boundaries and a Framework for Establishing Safe, Effective AI Tools. *JMIR Mental Health*, 12, e84854.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., ... & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, 22(3), e16235.

Tavory, T. (2024). Regulating AI in Mental Health: Ethics of Care Perspective. *JMIR MENTAL HEALTH*, 11. (WOS:001322536000001). <https://doi.org/10.2196/58493>

Tekin, Ş., & Delehanty, M. (2025). Beyond doomsday fears: why we need to consider the potential harms of AI psychotherapy. *The American Journal of Bioethics*, 26(2), 45-55.

Trothen, T. J. (2022). Replika: Spiritual enhancement technology?. *Religions*, 13(4), 275.

U.S. Food & Drug Administration. (2025, November 6). *Executive summary for the Digital Health Advisory Committee meeting: Generative artificial intelligence-enabled digital mental health medical devices* (FDA Doc. No. 189391). <https://www.fda.gov/media/189391/download>

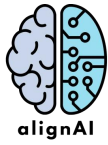
Vecchione, B., & Singh, R. (2025). Artificial intelligence is mental: Evaluating the role of large-language models in supporting mental health and well-being. *Big Data & Society*, 12(4), 20539517251383884.

Vilaza, G. N., & McCashin, D. (2021). Is the automation of digital mental health ethical? Applying an ethical framework to chatbots for cognitive behaviour therapy. *Frontiers in Digital Health*, 3, 689736.

Wells, K. (2023, June 9). *An eating disorders chatbot offered dieting advice, raising fears about AI in health*. NPR. <https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea>

Woebot Health. (2026). *Woebot: AI-powered mental health support*. Retrieved April 5, 2026, from <https://www.woebothealth.com>

Wrightson-Hester, A. R., Anderson, G., Dunstan, J., McEvoy, P. M., Sutton, C. J., Myers, B., ... & Mansell, W. (2023). An artificial therapist (manage your life online) to support the mental health of youth: co-design and case series. *JMIR Human Factors*, 10(1), e46849.



Wysa Ltd. (2026). *Wysa: AI-powered mental health support*. Retrieved April 5, 2026, from <https://www.wysa.com>

Xie, T., & Pentina, I. (2022). *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika*. Hawaii International Conference on System Sciences. <https://doi.org/10.24251/HICSS.2022.258>

Xu, Z., Lee, Y.-C., Stasiak, K., Warren, J., & Lottridge, D. (2025). The Digital Therapeutic Alliance With Mental Health Chatbots: Diary Study and Thematic Analysis. *JMIR Mental Health*, 12, e76642. <https://doi.org/10.2196/76642>

Yang, F., & Oshio, A. (2025). Using attachment theory to conceptualize and measure the experiences in human-AI relationships. *Current Psychology*, 44(11), 10658-10669.

Youn, S., & Jin, S. V. (2021). "In AI we trust?" The effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy." *Computers in Human Behavior*, 119, 106721.

Youper, Inc. (2026). *Youper: AI mental health*. Retrieved April 5, 2026, from <https://www.youper.ai>

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.

Zhang, A., & Rau, P. L. P. (2023). Tools or peers? Impacts of anthropomorphism level and social role on emotional attachment and disclosure tendency towards intelligent agents. *Computers in Human Behavior*, 138, 107415.

Zhang, X., Zayed, A., Rehn Hamrin, J., Güneysu, A., & Kuoppamäki, S. (2025). Exploring Body Image Awareness With a Large Language Model-Based Conversational Agent: Qualitative Study With Young Adults. *Journal of Medical Internet Research*, 27, e78829.