Patient Safety at Risk? Exploring Risks and Opportunities of AI Conversational Tools in Mental Health and Towards a Layered Safety Framework

Josephine Blonder

Dissertation written under the supervision of Professor Henrique Martins

Dissertation submitted in partial fulfilment of requirements for the MSc in Management Strategic Marketing, at the Universidade Católica Portuguesa, 05.01.2026

Abstract

The increasing use of AI-driven conversational tools in mental-health contexts has intensified debates on patient safety, ethical responsibility, and governance. While these systems promise improved accessibility and early support, they also introduce risks of psychological harm and insufficient safeguards, particularly for vulnerable users. Existing approaches often address these risks in isolation and fail to capture how safety failures emerge across interacting socio-technical layers.

This dissertation examines how patient safety in AI-driven conversational mental-health tools can be conceptualised and safeguarded through a layered framework. Drawing on patient-safety theory, particularly the Swiss Cheese Model, it adopts a qualitative, exploratory design to analyse how responsibilities, vulnerabilities, and safeguards interact across regulatory, technical, clinical, and user layers. Empirical insights were generated through semi-structured interviews with clinicians, industry experts, regulators, and patients.

The findings show that patient safety risks arise not from isolated failures but from the alignment of vulnerabilities across layers. Responsibilities are widely distributed yet insufficiently coordinated, interdependencies at interfaces are critical safety determinants, and reliance on user responsibility is structurally unreliable in mental-health contexts.

In summary, the study concludes that:

- Patient safety in conversational mental-health AI is a system-level outcome shaped by cross-layer interactions.
- Ethical responsibility becomes a source of risk when responsibilities are implicitly delegated.
- Governance is effective only when implemented as layered, cross-cutting mechanisms.

Title: *Patient Safety at Risk? Exploring Risks and Opportunities of AI Conversational Tools in Mental Health and Towards a Layered Safety Framework*

Author: Josephine Blonder

Keywords: patient safety; conversational AI; mental health; AI governance; socio-technical systems

Sumário Executivo

O uso crescente de ferramentas conversacionais baseadas em inteligência artificial na saúde mental intensificou os debates sobre segurança do paciente, responsabilidade ética e governação. Embora estes sistemas prometam melhorar a acessibilidade, também introduzem riscos de dano psicológico para utilizadores vulneráveis. Abordagens existentes tendem a tratar estes riscos de forma isolada, não captando a sua natureza sociotécnica.

Esta dissertação analisa como a segurança do paciente em ferramentas conversacionais de saúde mental baseadas em IA pode ser conceptualizada e salvaguardada através de uma estrutura em camadas. Com base no Modelo do Queijo Suíço, o estudo adota um desenho qualitativo e exploratório para analisar a interação entre responsabilidades, vulnerabilidades e salvaguardas nas camadas regulatória, técnica, clínica e do utilizador, com base em entrevistas a múltiplos stakeholders.

Os resultados indicam que os riscos de segurança decorrem do alinhamento de vulnerabilidades entre camadas, sobretudo em contextos de sofrimento psicológico. As responsabilidades estão distribuídas, mas pouco coordenadas, e a dependência da responsabilidade do utilizador revela-se inadequada.

Em síntese, conclui-se que:

- A segurança do paciente em IA conversacional é um resultado sistémico moldado por interações entre camadas.
- A responsabilidade ética torna-se um risco quando não é explicitamente coordenada.
- A governação é eficaz apenas quando implementada através de mecanismos em camadas.

Título: *A Segurança do Paciente em Risco? Exploração dos Riscos e Oportunidades das Ferramentas Conversacionais de IA na Saúde Mental e Rumo a uma Estrutura de Segurança em Camadas*

Autora: Josephine Blonder

Palavras-chave: segurança do paciente; IA conversacional; saúde mental; governação da IA; sistemas sociotécnicos

Table of Content

1. Introduction

In June 2025, the parents of a sixteen-year-old boy filed a lawsuit against OpenAI, alleging that their son's suicide had been influenced by interactions with a conversational AI system. While legal responsibility in such cases remains contested, the case drew global attention to the growing presence of AI-driven conversational tools in mental-health contexts and to the difficulties of determining how these systems relate to psychological vulnerability, patient/individual safety, ethics of using emotionally impacting-AI tools, and responsibility. More broadly, it highlighted the complexity of safeguarding users when emotionally responsive technologies are deployed in situations of distress.

AI-driven conversational tools are increasingly used in mental-health contexts to provide emotional support, psychoeducation, and guidance through text- and voice-based interfaces (Li et al., 2023; Maurya et al., 2024; Sarkar et al., 2023). Promoted as accessible, scalable, and stigma-reducing, these systems are frequently positioned as complements to, or partial substitutes for, traditional care, particularly in settings characterised by limited clinical capacity, high costs, and long waiting times (Balan et al., 2024; Kulkarni et al., 2024; van der Schyff et al., 2023). At the same time, their growing use by psychologically vulnerable individuals has raised pressing concerns about patient safety, responsibility allocation, and appropriate governance (Meadi et al., 2025; Sarkar et al., 2023; Wang et al., 2024).

Unlike many digital health technologies, conversational AI systems engage users through ongoing, emotionally responsive dialogue and increasingly aim to emulate empathic therapist–client interactions (Chu et al., 2024; Ingle, 2025; Mansoor et al., 2025). By simulating empathy, offering continuous interaction, and sometimes being framed as companions, they shape how distress is interpreted, managed, and escalated (Balcombe, 2023; Maurya et al., 2024). While these characteristics may lower barriers to help-seeking and self-disclosure (Lee et al., 2025; Sharma et al., 2022), they also introduce novel risks, including delayed escalation in crisis situations, harmful or misleading advice, emotional over-reliance, and ambiguity regarding the system's role and authority (Gamble, 2020; Meadi et al., 2025; Wang et al., 2024). Recent public debates and regulatory initiatives reflect these concerns, yet uncertainty remains about how safety failures emerge in practice and how responsibility should be distributed across designers, deployers, clinicians, and users (Balcombe, 2023; Miner et al., 2019; Sedlakova & Trachsel, 2022).

Academic and policy-oriented discussions address aspects of these challenges through digital mental-health research, AI governance, and ethical analysis. However, these perspectives often

remain fragmented. Safety is frequently treated as a matter of technical robustness, regulatory compliance, or user responsibility in isolation, rather than as an outcome emerging from interactions between regulatory, technical, clinical, and user layers. As a result, there is limited empirical understanding of how misalignments between these layers contribute to patient-safety risks in real-world settings, particularly in contexts of psychological vulnerability.

This thesis addresses this gap by examining how patient safety, particularly mental integrity and safety, in contexts of AI-driven conversational mental-health tools use, can be conceptualised and safeguarded through a layered socio-technical framework. The overarching research question asks how patient safety can be understood and protected across interacting layers, rather than within individual components. Four sub-questions guide the analysis, focusing on safety responsibilities, vulnerability formation, cross-layer interdependencies, and the role of safeguards and governance mechanisms.

The study adopts a qualitative, exploratory research design based on semi-structured interviews with clinicians, industry experts, regulators, and patients. By integrating patient-safety theory with stakeholder perspectives, the thesis aims to contribute a refined framework that supports more coherent, system-level approaches to safeguarding conversational AI in mental-health contexts.

The remainder of the thesis is structured as follows. Chapter 2 outlines the conceptual background and derives the research questions. Chapter 3 describes the methodology. Chapter 4 presents the empirical findings. Chapter 5 discusses these findings as well as its limitations, and Chapter 6 concludes by answering the research question.

2. Background

2.1 Context and Relevance: AI in Mental Health

This chapter provides the conceptual foundation for analysing patient safety in AI-driven conversational tools. It outlines how these systems function, the opportunities and risks they present, the regulatory and ethical conditions shaping their development, and the frameworks used to conceptualise safety in layered socio-technical environments. The aim is to establish the key constructs and systemic factors required to understand how vulnerabilities emerge, how responsibilities and safeguards intersect across stakeholder groups, and how these dynamics influence the safe use of conversational AI in mental-health contexts.

To do so, the chapter first maps AI applications in healthcare and mental health, drawing on evidence regarding uptake, benefits, risks, and the experiences of patients and clinicians (Abd-Alrazaq et al., 2019; Blease & Torous, 2023; Hipgrave et al., 2025; Sharif et al., 2025). It then introduces the regulatory and ethical landscape relevant to safety evaluation, before turning to patient-safety theory, where the Swiss Cheese Model serves as the primary organising lens (Reason, 2000; Wiegmann et al., 2022; Perneger, 2005). Building on these strands, Section 2.5 synthesises the current knowledge base to highlight what is known, what remains uncertain, and where existing frameworks fall short. The chapter concludes by linking these foundations to the methodological approach, which uses expert interviews with regulators, industry experts, clinicians, and patients to refine a tailored layered framework for safeguarding vulnerable users.

2.2 AI in Mental Health: Current Landscape

2.2.1 AI in Healthcare Broadly

Artificial intelligence has moved from experimental pilot projects to increasingly embedded use across health systems, shaping how care is planned, delivered, and governed. Recent evidence from the World Health Organization Regional Office for Europe's region-wide assessment of artificial intelligence in health care, based on a 2024–2025 survey of 50 Member States, characterises this shift as a system-level transformation encompassing clinical applications, governance arrangements, workforce preparedness, and health data infrastructures (WHO Europe, 2025). The report indicates that AI is no longer confined to backend clinical decision support but is increasingly deployed in patient-facing contexts: approximately half of participating Member States reported the use of AI-driven chatbots for patient support, signalling that conversational interfaces are becoming part of routine health-system practice (WHO Europe, 2025). At the same time, the assessment highlights significant gaps in

institutional readiness. Only a minority of Member States reported having a health-specific AI strategy in place, and legal and regulatory uncertainty was identified as the most frequently cited barrier to adoption, reflecting ongoing ambiguity around accountability, oversight, and safe implementation at scale (WHO Europe, 2025). Within this rapidly evolving landscape, the broad diffusion of AI in health care provides the technical and institutional foundation for its expansion into mental health, while simultaneously foregrounding a central challenge for this thesis: as AI becomes more accessible through conversational modalities, questions of patient safety, responsibility allocation, and enforceable safeguards become increasingly salient, particularly in settings where vulnerability and risk escalation are difficult to detect, classify, and manage. This development mirrors broader trends in digital mental-health interventions, where AI-based tools are increasingly embedded alongside teletherapy, monitoring apps, and blended care models, raising new coordination and safety challenges (Löchner et al., 2025).

2.2.2 Conversational AI in Mental Health

Conversational AI systems simulate dialogue with users through natural-language processing and vary widely in sophistication. Early mental-health chatbots such as Woebot and Wysa used structured cognitive-behavioural-therapy (CBT) scripts, whereas newer systems rely on large language models such as GPT-4 to generate adaptive, context-aware responses (Vaidyam et al., 2019; Kalam, 2024). This shift from scripted to generative interaction expands flexibility but also unpredictability, introducing new demands for safety governance and behavioural constraints (Chakraborty et al., 2023). These systems are typically accessed through mobile apps or web platforms and marketed as low-cost, scalable companions for stress management, mood tracking, or emotional support (Abd-Alrazaq et al., 2019). Consumer-facing tools like Replika further blur boundaries between support, relationship simulation, and therapy, raising concerns about dependency, attachment dynamics, and insufficient clinical oversight (Parks et al., 2025). Recent systematic reviews show that conversational AI is moving from experimental deployments to operational use in both self-help and clinical settings (Ali et al., 2025; Dehbozorgi et al., 2025). Qualitative studies report that many users experience these tools as emotionally supportive, accessible, and easy to approach (Siddals et al., 2024). However, concerns persist about the reliability of generated advice and the lack of explicit boundaries distinguishing emotional support from therapeutic guidance. These developments highlight the importance of robust safety protocols and carefully defined integration pathways.

2.2.3 Benefits: Accessibility, Scalability, Anonymity

The most cited benefit of conversational AI is accessibility. Conversational AI is continuously available, reduces barriers such as distance, cost, stigma, and wait times, and can support underserved populations (WHO, 2021; Olawade et al., 2024). Its scalability allows support to be offered to large populations at low marginal cost, which is particularly valuable given global shortages of mental-health professionals (OECD, 2019; Dehbozorgi et al., 2025).

Apart from accessibility, anonymity is another widely reported advantage. Users often disclose personal or sensitive experiences more freely to AI systems than in face-to-face settings, perceiving them as non-judgmental and emotionally neutral (Vaidyam et al., 2019; Abd-Alrazaq et al., 2019). Narrative studies highlight that many individuals describe these interactions as comforting and easy to approach, especially during periods of distress or social withdrawal (Siddals et al., 2024).

Another important benefit of conversational AI is its potential to contribute to prevention. This aligns with scoping evidence showing that AI-driven mental-health tools are increasingly used across screening, support, monitoring, and prevention functions (Ni & Jia, 2025). Linguistic-analysis models have demonstrated potential for identifying early indicators of depression, stress, or suicide risk, allowing timely intervention or triage (Rollwage et al., 2023).

These benefits suggest strong potential for enhancing reach, early detection, and equity across mental-health systems.

2.2.4 Risks: Misinformation, Empathy Deficits, and Privacy Concerns

Despite their benefits, conversational AI tools present significant risks. Generative models can produce inaccurate or misleading information ("hallucinations"), which is particularly dangerous when users seek support in moments of crisis (Kalam, 2024; Parks et al., 2025).

The absence of genuine empathy limits therapeutic value, as emotional attunement and trust are central to care (Blease & Torous, 2023; Brown & Halpern, 2021). Recent audits show inconsistent empathy and demographic bias in LLM-generated responses, raising concerns about fairness and reliability (Omar et al., 2024).

Privacy risks further complicate deployment: many systems collect sensitive mental-health data, raising questions about informed consent and compliance with data protection frameworks such as the GDPR (Busch et al., 2024). The rapid public adoption of commercial generative-AI chatbots, often in unregulated settings, is outpacing systematic safety validation and oversight, creating emerging risks for users and society at scale (Migliorini, 2024). High-profile lawsuits

following suicide cases allegedly linked to ChatGPT illustrate how technical, ethical, and regulatory weaknesses can converge with severe consequences (CNN, 2025; TIME, 2025).

Recent evaluations also show that LLM-based systems display fluctuating or misattuned emotional responses across contexts and user profiles. Some interactions appear supportive, while others shift in tone or produce emotionally incongruent messages (Varastehnezhad et al., 2025). In mental-health contexts, such variability may be interpreted as meaningful reassurance, especially among distressed users. Recent evaluations further indicate that LLM-based conversational systems can display fluctuating or contextually misaligned emotional responses across users and interaction settings. In mental-health contexts, such variability may be interpreted by distressed users as meaningful reassurance or therapeutic attunement. A recent systematic review by Guo et al. (2024) highlights that while large language models are increasingly applied to mental-health support tasks, their empathic responses remain inconsistent and insufficiently grounded in clinical judgment, raising concerns about emotional over-reliance and delayed escalation in high-risk situations. These behavioural limitations underscore the need for robust safeguards when conversational agents interact with psychologically vulnerable populations (Choudhury, 2020). Recent analyses further emphasise that such harms often remain under-recognised because they arise from interactional and governance failures rather than explicit technical errors (Denecke et al., 2025).

2.2.5 Perspectives: Patients and Clinicians

User perspectives reveal both appreciation and concern. Patients often value the immediacy, anonymity, and non-judgmental stance of chatbots, describing them as approachable and emotionally supportive (Sharif et al., 2025; Siddals et al., 2024). At the same time, qualitative studies note worries about emotional over-reliance, misattuned responses, and the absence of genuine empathy. Users also report uncertainty about when AI-generated guidance is appropriate and how to distinguish emotional support from therapeutic judgement (Brown & Halpern, 2021).

Clinicians express a similarly mixed view. They recognise the usefulness of conversational AI for triage, psychoeducation, and symptom monitoring (Hipgrave et al., 2025; MacNeill et al., 2024), yet caution against unsupervised deployment due to risks of misinterpretation, missed crisis cues, and unrealistic expectations of AI competence. Evidence suggests that clinician trust correlates strongly with digital literacy and the clarity of regulatory guidance (Baek et al., 2025).

Taken together, patient and clinician perspectives point toward a shared conclusion: conversational AI offers meaningful potential but requires structured responsibilities, clear boundaries, and robust safety governance to support its responsible use (Goktas & Grzybowski, 2025)

2.3 Regulatory and Ethical Context

2.3.1 European Frameworks

Within the European Union, AI in healthcare is governed by several overlapping regulatory instruments that seek to balance innovation with patient safety. The EU Artificial Intelligence Act (European Commission, 2024) classifies most health-related AI systems, including conversational mental-health applications, as high-risk. These systems are subject to conformity assessment, human oversight, and post-market monitoring requirements. This classification reflects the potential for psychological and informational harm, particularly when such tools interact directly with vulnerable users.

The Medical Device Regulation (MDR; EU 2017/745) applies to AI systems that qualify as medical devices, including diagnostic tools and digital therapeutics. Developers must demonstrate clinical safety, efficacy, and quality management in line with established medical-device standards (TÜV AI Lab, 2023). However, many consumer-facing chatbots, such as Replika or Woebot, do not claim therapeutic intent and therefore fall outside the MDR. This creates a regulatory gap in which systems can simulate therapeutic dialogue without corresponding clinical accountability (Busch et al., 2024).

The GDPR further governs the processing of personal and sensitive data, requiring explicit consent, transparency, purpose limitation, and safeguards against secondary use (Pinsent Masons, 2024). Conversational AI tools often collect highly emotional disclosures, raising concerns when such data are stored insecurely, reused for model training, or shared with third parties without meaningful consent (Meszaros et al., 2022).

Recent scholarship also points to structural challenges that existing regulation does not fully address. Responsibility in modern AI systems is distributed across model providers, application developers, platform hosts, and third-party integrators. This diffusion complicates accountability when harm occurs and conflicts with regulatory frameworks that still assume a single manufacturer with end-to-end control (Widder & Nafus, 2022). Such ambiguity is particularly problematic in mental-health contexts, where fragmented oversight can hinder incident investigation and prevention.

A further limitation concerns psychological safety. While high-risk classification acknowledges potential harm, neither the AI Act nor the MDR defines psychological harm in operational terms. As a result, regulation prioritises technical robustness and data governance while offering limited guidance on risks such as distress-driven suggestibility, pseudo-empathy, or emotional dependency (Huang et al., 2025; Chu et al., 2025).

Together, these instruments form a layered but fragmented regulatory regime that continues to struggle to keep pace with rapidly evolving conversational mental-health technologies.

2.3.2 Global Frameworks

Globally, several organisations have articulated high-level principles for AI governance in healthcare. The World Health Organization's Global Strategy on Digital Health 2020–2025 highlights AI's potential to expand access while stressing accountability, transparency, equity, and meaningful human oversight (WHO, 2025). Notably, WHO explicitly includes cognitive and emotional risks within patient safety, which is particularly relevant for mental-health applications.

The OECD Principles on Artificial Intelligence promote human-centred values, fairness, explainability, robustness, and accountability and have shaped international consensus despite their non-binding nature (OECD, 2019; Floridi, 2024). In the United States, FDA guidance on adaptive AI medical devices focuses on transparency and post-market monitoring but does not yet address conversational agents specifically (Weiner et al., 2025). The lack of harmonised international standards complicates compliance and leaves psychological safety insufficiently specified.

Across jurisdictions, regulation continues to lag technological development. This pacing problem creates gaps around emerging risks such as emotionally adaptive interaction, and distress escalation, which are especially critical in mental-health contexts (Bengio et al., 2024).

Ethical guidance further informs responsible design. Core bioethical principles non-maleficence, beneficence, autonomy, and justice remain central and require conversational systems to avoid harm, support well-being, ensure transparency and consent, and mitigate bias (Beauchamp & Childress, 2019). Recent work emphasises that ethics and safety are inseparable, particularly where psychological integrity may be undermined by over-reliance, false reassurance, or misleading advice (Blease & Torous, 2023; Floridi, 2024; Weiner et al., 2025). This has led to calls for operational safeguards that translate ethical principles into concrete design and governance measures (Meadi et al., 2025).

Empirical evidence suggests that existing safeguards remain insufficient. Studies show that large language models in mental-health contexts can reproduce stigma, provide inappropriate guidance, or display biased emotional responses, underscoring the limits of principle-based ethics without structured oversight (Moore et al., 2025).

Overall, global legal and ethical frameworks establish important boundaries for responsible AI use but remain fragmented and unevenly enforced, leaving many conversational mental-health systems outside formal oversight and users exposed to psychological and informational risks.

2.4 Patient Safety and Theoretical Frameworks

2.4.1 Defining Patient Safety

Patient safety traditionally refers to the prevention of harm within healthcare systems. The World Health Organization (WHO, 2025) defines it as the absence of preventable harm and the reduction of unnecessary risk. Early safety science centred on physical harms such as surgical errors, medication mistakes, and hospital-acquired infections (Van Gerven et al., 2016).

More recently, patient safety has expanded to include psychological safety and mental integrity (Blease & Torous, 2023). In mental-health contexts, harm may arise not from procedures but from misleading information, distress amplification, inappropriate advice, or breaches of emotional privacy.

2.4.2 The Swiss Cheese Model

The Swiss Cheese Model (Reason, 2000) remains one of the most influential frameworks in patient safety. It conceptualises safety as a system of multiple defensive layers, each containing weaknesses or "holes." Harm occurs when these weaknesses align, enabling hazards to pass through all layers (Wiegmann et al., 2022).

In healthcare, this model explains how overlapping safeguards, such as verification steps, cross-checks, and monitoring, create redundancy that prevents errors. The model also emphasises a systems perspective: failures emerge not from a single mistake but from interactions among structural, technical, and human factors (Perneger, 2005).

Recent critiques note that vulnerabilities in digital and socio-technical systems are rarely random; instead, they stem from organisational pressures, design trade-offs, or biased data pipelines (Denecke et al., 2023). Despite these limitations, the Swiss Cheese Model remains a powerful heuristic for conversational AI because it captures how failures can originate at

different layers, model training, regulatory gaps, clinician oversight, or user behaviour and align to produce harm.

2.4.3 Application to AI in Mental Health

Applying the Swiss Cheese Model to AI conversational tools requires identifying where layers of defence exist and where vulnerabilities emerge across technical, regulatory, clinical, and user domains.

Possible Layers of Defence

- Developer safeguards: model training standards, bias audits, constrained generation, and guardrails to prevent unsafe outputs (Jimini Health, 2025).
- Regulatory oversight: compliance with the EU AI Act, MDR, GDPR, and post-market monitoring obligations (Busch et al., 2024).
- Clinical integration: ensuring AI supports, rather than replaces, clinician judgement, especially in high-risk contexts (Golden & Aboujaoude, 2024).
- User education: improving digital literacy, expectation-setting, and understanding of limitations (Gabriel, 2024).
- Escalation protocols: reliable mechanisms to detect suicidal ideation or acute distress and direct users to human support (Iorfino et al., 2017).

Common "Holes" in These Layers

- Biased or incomplete training data that distort outputs or miss key crisis signals.
- Inconsistently triggered or overly narrow escalation pathways, particularly for short or fragmented suicidal expressions.
- Over-reliance by vulnerable users who anthropomorphise AI or misconstrue pseudo-empathy as genuine therapeutic competence.
- Weak or absent regulation when tools fall outside MDR classification, creating accountability gaps.
- Low digital literacy and unclear data-handling practices that undermine informed consent.

Emerging research on technological folie à deux demonstrates how conversational agents and users can mutually amplify maladaptive patterns, reinforcing distress rather than alleviating it

(Dohnányi et al., 2025). A single case illustrates how vulnerabilities can align: a user experiencing suicidal thoughts may engage with an unregulated chatbot lacking escalation protocols, while clinicians remain unaware of its use, and the user lacks the literacy to recognise its limits.

Eliminating all vulnerabilities is neither feasible nor desirable; overly rigid safeguards could reduce accessibility and undermine early intervention potential. The challenge lies in designing proportionate, layered safety architectures that reduce the likelihood of systemic failures while preserving access for users who benefit from conversational support.

2.4.4 Other Relevant Theoretical Lenses

While the Swiss Cheese Model provides the central analytical structure of this study, additional frameworks are used to complement it by offering more granular criteria for assessing risk in AI-based mental-health systems. The Framework for AI Tool Assessment in Mental Health (FAITA) proposes structured evaluation criteria covering clinical validity, data protection, ethical robustness, and suitability for specific patient populations (Golden & Aboujaoude, 2024). The Digital Therapeutics Risk Assessment Canvas extends this perspective by mapping risks across design, implementation, and user-interaction phases, supporting the early identification of weaknesses in complex socio-technical systems (Denecke et al., 2023). Together, these frameworks operationalise the Swiss Cheese Model by enabling systematic assessment of cognitive, emotional, behavioural, and relational risks that may emerge when users interact with AI systems during periods of vulnerability.

The Swiss Cheese Model was selected as the primary analytical lens because it conceptualises harm as the result of aligned weaknesses across multiple layers rather than isolated failures within a single component (Reason, 2000; Wiegmann et al., 2021). This systemic perspective is particularly suited to conversational AI in mental health, where risks emerge at the intersection of regulatory oversight, technical design, clinical practice, and user vulnerability, and where socio-technical factors combine to produce harm rather than discrete “errors” alone (Cummings, 2024). Unlike principle-based ethical frameworks or governance-focused models, which tend to address risks in isolation or at a single level of abstraction, the Swiss Cheese Model explicitly captures interaction effects, handovers, and responsibility gaps across socio-technical layers (Perneger, 2005; Seshia et al., 2017; Shabani et al., 2023).

Alternative approaches were considered but not adopted as the central framework. Ethical principles such as beneficence and autonomy provide important normative guidance but offer limited explanatory power for analysing how safety failures materialise in practice, particularly

in complex systems where latent conditions and active failures interact over time (Reason, 2000). Regulatory and accountability frameworks primarily address upstream compliance and institutional responsibility, while human–computer interaction and trust models focus on user perception without capturing system-level safety dynamics in full (Larouzée & Le Coze, 2020; Perneger, 2005). The Swiss Cheese Model was therefore chosen because it uniquely allows patient safety in conversational mental-health AI to be analysed as a layered, dynamic system, while remaining compatible with complementary frameworks that provide domain-specific evaluation criteria (Cummings, 2024; Shabani et al., 2023).

2.5 Synthesis and Gap Identification

The literature on conversational AI in mental health highlights both substantial potential and persistent safety concerns (Meadi et al., 2025). A growing body of empirical and review-based research demonstrates that conversational agents can expand access to support, reduce stigma, and facilitate early intervention by offering low-cost, continuous, and often anonymous assistance that users may perceive as less intimidating than face-to-face care (Kaywan et al., 2023; Li et al., 2023; Mansoor et al., 2025). These systems are increasingly explored for psychoeducation, symptom monitoring, and the detection of early warning signs, positioning them as a promising complement to traditional services in contexts characterised by workforce shortages and access barriers (Golden & Aboujaoude, 2024; Meadi et al., 2025). Importantly, the deployment of conversational AI in mental-health contexts builds upon long-standing regulatory, ethical, and patient-safety principles developed in traditional healthcare, where compliance, accountability, and harm prevention are central to maintaining quality of care and public trust (Mokeke et al., 2024; Sharma et al., 2023).

Building on this foundation, several recent frameworks and reviews propose structured approaches to evaluating the technical, clinical, and ethical risks associated with AI and generative AI in mental health. Scoping and systematic reviews consistently identify recurring concerns related to privacy, bias, safety, accountability, and simulated empathy, and increasingly argue for multidimensional or layered accountability models to guide responsible deployment (Balcombe, 2023; Lei, 2025; Meadi et al., 2025; Miner et al., 2019; Ni & Jia, 2025; Wang et al., 2024). Patient-safety frameworks such as the Swiss Cheese Model further contribute a systemic lens by conceptualising harm as the alignment of weaknesses across interacting layers, an approach that has been adapted to digital health and AI-enabled interventions through tools such as FAITA-MH and the Digital Therapeutics Risk Assessment Canvas (Denecke et al., 2023; Golden & Aboujaoude, 2024; Reason, 2000; Wiegmann et al.,

2022). Together, these contributions demonstrate growing recognition that AI-related risks cannot be assessed in isolation and must be understood within broader socio-technical systems. Despite these advances, important gaps remain in both research and practice. Existing approaches lack a unified, layered framework that integrates regulatory, ethical, technical, and psychological safeguards in a coherent manner tailored to conversational AI in mental-health contexts. Empirical research rarely links lived psychological vulnerability and real-world user experience to system-level protections and governance arrangements, leaving a disconnect between how risk is experienced by users and how safety is designed upstream. Moreover, there is limited guidance on operational mechanisms that translate high-level regulatory and ethical principles into concrete, measurable safeguards that can be implemented, coordinated, and evaluated by developers, clinicians, and deployers in practice.

Building on these gaps, this thesis aims to examine how patient safety in AI-driven conversational mental-health tools can be conceptualised and safeguarded through a layered framework. Drawing on different qualitative perspectives, it approaches patient safety not as a matter of technical accuracy, regulatory compliance, or individual user responsibility in isolation. Instead, the study adopts a system-level perspective, treating safety as an emergent property of a layered socio-technical arrangement shaped by the interaction, alignment, and mutual compensation of regulatory, technical, clinical, and user layers.

2.6 Research Questions, Sub-Questions and Hypotheses

Building on the literature on patient safety, socio-technical systems, and AI governance in mental health, this study conceptualises patient safety as a layered phenomenon shaped by the interaction of regulatory, technical, clinical, and user layers. Rather than testing causal relationships, the study adopts a qualitative and exploratory approach in which theory-informed expectations are examined and refined through empirical analysis. Accordingly, the following research questions and hypotheses guide the study. The overarching research question is:

RQ: *How can patient safety in AI-driven mental-health tools, particularly conversational agents, be conceptualised and safeguarded through a layered framework?*

To address this question, the analysis is structured around four sub-questions:

- **SQ1 (Responsibilities & Controls):**
  What safety responsibilities and controls exist, or should exist, at each layer: Clinicians (C), Industry Experts (IE), Patients (P), and Regulators (R)?
- **SQ2 (Vulnerabilities):**
  Where do vulnerabilities ("holes") most commonly arise within each layer, and how do they align across layers to create risk?
- **SQ3 (Cross-Layer Coordination / Interdependencies):**
  What interdependencies at the interfaces between layers (e.g. Clinicians (C), Industry Experts (IE), Patients (P), and Regulators (R)) most influence safety outcomes?
- **SQ4 (Safeguards & Governance):**
  Which practical safeguards and governance mechanisms could strengthen patient safety while maintaining accessibility and innovation?

Based on the existing literature, one hypothesis is formulated for each sub-question. These hypotheses represent theory-driven expectations that orient the analysis and are assessed and refined in the Discussion chapter.

- **H1:** Patient safety responsibilities are distributed across layers but insufficiently aligned.
- **H2:** Patient safety risks arise primarily from the alignment of vulnerabilities across layers.
- **H3:** Safety outcomes are strongly shaped by interdependencies at interfaces between layers.
- **H4:** Safeguards enhance patient safety only when they operate across layers rather than in isolation.

3. Methodology

3.1 Research Design

This dissertation adopted a qualitative, exploratory research design to examine how patient safety can be conceptualised and safeguarded in the deployment of AI-driven conversational tools in mental-health contexts. Rather than testing predefined hypotheses, the study aimed to develop and refine a layered patient-safety framework, informed by the Swiss Cheese Model, through the systematic synthesis of expert perspectives across the mental-health AI ecosystem. A qualitative approach was considered appropriate given the emergent, socio-technical, and context-dependent nature of conversational AI in mental health. While existing literature documents both potential benefits and risks, it provides limited guidance on how abstract safety principles can be operationalised across regulatory, technical, clinical, and user-facing layers. Expert-based qualitative inquiry therefore enabled the translation of theoretical safety concepts, responsibilities, and failure points as they manifest in practice.

Empirical data were generated through semi-structured interviews with participants drawn from four stakeholder groups, selected a priori based on their distinct roles within the patient-safety system:

Regulators and legal experts (R) with experience in EU digital regulation, including the AI Act, the Medical Device Regulation (MDR), and the General Data Protection Regulation (GDPR), representing the legal and governance layer. Industry experts and developers (IE) involved in the design, validation, or strategic deployment of AI-based chatbots and digital therapeutics, representing the technical and organisational layer. Clinicians (C), including psychiatrists, psychologists, and clinically adjacent professionals, representing the clinical decision-making and care-delivery layer. Patients (P) with lived experience of psychological vulnerability, included selectively to capture how system-level design and governance decisions are experienced at the user interface.

This stakeholder configuration reflects the logic of the Swiss Cheese Model, in which patient safety emerges from the interaction of multiple defensive layers rather than from any single actor. Each group contributed a distinct safety perspective, enabling the identification of safeguards, role boundaries, and potential misalignments between layers. Patient perspectives were incorporated to ground expert assessments in lived experience, while remaining analytically complementary rather than dominant, given the study's focus on systemic safety design. The research design maintained continuous alignment between background literature, research questions, and empirical data collection. Insights from scholarship on patient safety, psychological integrity, and regulatory oversight informed the interview guides, while thematic

analysis was used to map empirical findings back onto the layered framework. This iterative alignment supported the development of a coherent, practice-oriented guidelines that balance safety, accessibility, and innovation in digital mental-health care.

3.2 Sampling Strategy and Recruitment

The study employed a purposive, non-probabilistic sampling strategy aimed at capturing heterogeneous perspectives across the mental-health AI ecosystem. Sampling did not seek representativeness or statistical generalisation; instead, analytical relevance was derived from diversity in professional roles, institutional positions, and proximity to patient risk.

Participants were recruited through a multi-step process combining targeted professional outreach and referral-based sampling. The primary recruitment channel was LinkedIn, which was used to identify potential participants based on predefined criteria, including stakeholder group affiliation, professional involvement with mental-health AI, regulatory or clinical relevance, and geographic focus on Europe and comparable regulatory contexts. Approximately 35 individuals were contacted via LinkedIn, followed by 10 additional contacts via email where direct outreach was feasible. All email contacts received follow-up messages. This resulted in approximately seven responses from LinkedIn outreach and two responses from email contact. Additional participants were recruited through existing professional networks and referrals from interviewed experts.

This recruitment process reflects the practical challenges of accessing specialised experts in a rapidly evolving and highly regulated field. While response rates varied across stakeholder groups, sustained outreach efforts ensured sufficient coverage of relevant perspectives. The resulting sample comprised 14 interviews across four stakeholder groups and provided the heterogeneity required for qualitative, theory-informed analysis (Appendix A; Figure 1).

3.3 Data Collection and Interview Design

Data were collected through semi-structured interviews conducted between October and December 2025. Interviews were conducted synchronously via video call. In a limited number of cases, where scheduling constraints prevented live interviews, participants provided written responses via email. These responses followed the same interview guide and were treated as equivalent qualitative data for analytical purposes.

Interview guides were structured around the study's research sub-questions and reflected the layered logic of the patient-safety framework. Core thematic sections were consistent across all interviews to ensure comparability; however, individual questions were adapted to participants'

professional backgrounds and areas of expertise. For example, regulatory participants were asked to reflect on legal standards, accountability, and governance mechanisms, while clinicians were asked to focus on therapeutic risk, escalation, and clinical responsibility. Industry experts were prompted to discuss system design, validation, and boundary-setting, and patients were asked to reflect on lived experience and perceived safety at the user interface.

This adaptive interview design allowed for both structural consistency and contextual relevance. By tailoring question phrasing to participants' roles, interviews elicited domain-specific insights while maintaining alignment with the overarching research questions. This approach supported cross-stakeholder comparison without imposing artificial uniformity across fundamentally different perspectives.

Interviewees

ID	Stakeholder Group	Professional Role	Professional Experience	Country	Age Group	Gender
C01	Clinicians (C)	Psychologist & Psychological Psychotherapist in Training (acute psychiatry)	Clinical experience in acute psychiatry and crisis ward; work with suicidality, psychosis, schizoaffective disorders	Germany	30–35	female
C02	Clinicians (C)	Coach (Systemic Consultation; non-clinical)	Professional experience in systemic coaching with vulnerable groups; online coaching formats; work with young adults in stressful or transitional situations	Germany	25–30	female
C03	Clinicians (C)	Physician & Medical AI Researcher	Clinical care experience combined with interdisciplinary research on AI systems in healthcare	Germany	30-35	female
C04	Clinicians (C)	Licensed Psychological Psychotherapist	~20+ years of clinical experience in outpatient psychotherapy	Germany	Not disclosed	female
IE01	Industry Experts (IE)	Chief Commercial Officer, Digital Therapeutics	Senior industry executive; 20+ years in market access, commercialisation, and development of regulated digital therapeutics	United States	55–60	male
IE02	Industry Experts (IE)	Developer / Technical Expert (AI/ML)	Early-career technical expert; several years of academic experience in AI/ML, NLP, bias & fairness research	United Kingdom	25-30	female
IE03	Industry Experts (IE)	Researcher, LLM Behaviour & Safety	3+ years in LLM alignment, behaviour shaping, and safety research at a frontier model lab	United Kingdom	25-30	male
IE04	Industry Experts (IE)	Senior Strategist, Digital Therapeutics	7+ years in mental-health digital health (UX, clinical product development, regulatory coordination)	Germany	35-40	female
P01	Patients (P)	Student / Young Adult (mental-health service user)	Lived experience with depression, mood instability, and ongoing psychotherapy; repeated use of AI chatbots during vulnerable phases	Germany	20-25	female
P02	Patients (P)	Young Adult (mental-health service user)	Lived experience with hypochondria, panic attacks, and health anxiety; frequent use of AI chatbots for reassurance and panic regulation	Germany	25-30	female
P03	Patients (P)	Young Adult (mental-health service user)	Lived experience with low mood, rumination, loneliness; use of AI chatbots for emotional regulation and cognitive structuring	Portugal	25-30	male
R01	Regulators (R)	Legal expert in EU AI governance & digital regulation	Lawyer specialising in EU digital regulation & AI governance	Germany	45-50	female
R02	Regulators (R)	Legal expert in EU AI governance & digital regulation	Early-career legal expert; completed legal studies and LL.M.; academic and regulatory-related experience in AI governance, GDPR, and high-risk technology oversight	United Kindom	25-30	female
R03	Regulators (R)	Legal scholar specialising in EU law, patients' rights & AI regulation	Early- to mid-career academic; several years of research experience in EU regulation, mental-health law, and technology governance	Italy	25-30	male

Table 1: Overview of Interviewees across Stakeholder groups

3.4 Transcription and Translation

Transcription followed a clean verbatim approach, preserving participants' wording and intended meaning while omitting filler words and non-substantive utterances that did not affect semantic content. This ensured analytical accuracy while maintaining readability.

Interviews were conducted in both English and German. For interviews conducted in German, original German transcripts were retained and translated into English for the purposes of analysis and reporting. Translation was supported by AI-assisted tools and subsequently reviewed by the researcher, a native German speaker, to ensure semantic equivalence and contextual accuracy. Any minor linguistic smoothing was limited to clarification of meaning and did not alter substantive content. All interview transcripts are provided in the Appendix D. This approach was chosen to preserve readability while ensuring transparency and traceability of the empirical material. Within the transcripts, interview questions are clearly structured into thematic sections corresponding to the study's research sub-questions. This sectioning reflects the interview guide structure and facilitates traceability between the interview material, the results tables, and the thematic analysis presented in the Results chapter.

3.5 Data Analysis

Interview data were analysed using a thematic analysis approach informed by Braun and Clarke (Braun & Clarke, 2006). Given the qualitative scope of the study and the moderate number of interviews (n = 14), analysis was conducted manually to allow close engagement with the data and contextual recall of individual interviews.

As a first step, all interview responses were organised in a structured extraction matrix aligned with the interview guide. For each interview question, condensed summaries of participants' responses were documented per interviewee, preserving key formulations and positions. This question-by-question organisation enabled systematic comparison across participants and stakeholder groups while maintaining analytical traceability (Appendix B; Figure 2-4).

In a second step, recurring patterns across interviews were identified and iteratively grouped into higher-order themes. Themes were developed inductively and refined through repeated comparison across stakeholder groups, with attention to convergence, divergence, and tension between perspectives. Where relevant, the relative prominence of themes was described using qualitative frequency indicators, rather than statistical measures (Appendix C, Figure 5).

For each theme, one to two illustrative quotes were selected to exemplify how participants articulated the issue in their own words. Quotes were attributed using anonymised interview codes (e.g., C03, IE02) and were used exclusively in the Results section to support descriptive reporting.

3.6 Ethical Considerations and Confidentiality

All participants were informed about the purpose of the study, the voluntary nature of participation, and their right to withdraw at any time. Informed consent was obtained prior to data collection. To ensure confidentiality, all interviews were anonymised and reported using coded identifiers.

For selected participants, additional confidentiality assurances were provided through non-disclosure agreements (NDAs), particularly where professional roles involved sensitive regulatory, clinical, or commercial information. Interview data were handled exclusively by the researcher and used solely for academic purposes within the scope of this thesis. All identifying information was removed during transcription, and transcripts were securely stored and deleted after completion of the research process.

4. Results

This chapter presents the findings from the qualitative interviews conducted with clinicians, regulators, industry experts, and patients. The results are organised by the four thematic areas guiding the study: Responsibilities & Controls, Vulnerabilities, Cross-Layer Coordination (Interdependencies), and Safeguards & Governance. Within each thematic area, the perspectives of each stakeholder group are reported separately. The analysis identifies themes that emerged across participants' accounts, based strictly on what interviewees described in relation to each topic.

4.1 Responsibility and Controls

The interview data provided insights into how different stakeholders described responsibilities and controls in the context of AI-based mental health tools. Across clinicians, industry experts, regulators, and patients, participants outlined how roles were distributed and which actors they viewed as carrying primary obligations for safety. The following paragraphs summarise these accounts for each stakeholder group.

Clinicians:

Across the interviews, clinicians referred to several themes including distributed responsibility, clear limits and transparency, and misaligned expectations about AI use. They described responsibility as divided across regulatory, development, clinical, and user roles, and noted that clinicians should decide whether a tool is suitable for a given therapeutic context, while patients should not rely on AI for diagnostic purposes. C01 emphasised this point by stating that "patients should not use AI for diagnoses or as a substitute for therapy." Interviewees also stressed that platforms need to explain their capabilities and restrictions in an accessible way, with C02 noting that systems "must be transparent, provide clear warnings, and communicate where their limits are." Clinicians described moments in which users assumed the system could perform tasks beyond its scope or used it without professional input, and C04 highlighted concerns about patient vulnerability, stating that patients "are often vulnerable; we cannot expect them to carry the main safety burden.

Industry Experts:

Across the interviews, industry experts referred to several themes, including defined roles across stakeholders, boundary-setting and transparency, and responsibility limits for developers. They described regulators as setting standards and ensuring safety, data protection,

and performance requirements, while developers carry out these requirements through quality management, evidence generation, and ongoing monitoring, with IE01 noting that "each group has a clear lane." Experts described the importance of avoiding implications of clinical capability and emphasised transparency, as IE03 explained that preventing the "illusion of clinical expertise" requires constrained generation, explicit disclaimers, guardrails, and attention to tone, which "can mislead people" even when content is safe. Interviewees also spoke about responsibility limits, with IE03 stating that developers cannot be responsible for how downstream applications configure a model once it is integrated into multiple interfaces, and IE04 describing that product teams cannot compensate for regulatory gaps or write clinical protocols.

Regulators:

Across the interviews, regulators referred to several themes including setting standards and normative requirements, operational responsibilities for developers, clinical responsibilities in implementation, and limited responsibilities for end users. They described regulators as defining rules and minimum standards for safety, data protection, and high-risk classification, with R03 noting that rights such as privacy and informed consent must be "operationalised." Interviewees stated that developers must translate these requirements into practice through risk management, data governance, testing, documentation, and ongoing monitoring, and R03 added that developers should not rely on the "black box" argument when limitations impact users. Regulators also described clinicians as responsible for integrating AI in ways that preserve clinical judgment and autonomy, with R01 explaining that healthcare providers "should not outsource all decision-making to the system, especially when users are in a psychologically fragile position." End users were described as holding only limited responsibility because of their vulnerability, and R02 noted that "too much responsibility cannot be placed on them," while R03 emphasised their rights to be informed, to understand system capabilities, and to refuse AI without losing access to care.

Patients:

Across the interviews, patients referred to several themes, including limited user responsibility, developer and regulator duty, and need for clear crisis guidance. They consistently stated that users should not carry the main safety responsibility when distressed, with P01 noting that "you can't always judge what's safe or healthy" and P02 explaining that "when you're anxious or in a panic, you're not thinking clearly enough to judge what's safe." Interviewees described

developers and regulators as the actors who should set limits, define what is allowed, and ensure robust protections, while clinicians were viewed as important because they understand crisis needs. Patients also mentioned situations where they would have wanted clearer direction toward human help during suicidal thoughts, intense panic, or emotional overwhelm, with P03 noting that "clearer signals or boundaries would have helped me step back and involve a real person sooner."

4.2 Vulnerabilities

The interviews highlighted a range of vulnerabilities that participants associated with AI-based mental-health tools. Across clinicians, industry experts, regulators, and patients, interviewees described situations in which systems, users, or contextual factors increased the likelihood of unsafe or unreliable interactions. The following paragraphs summarise how each stakeholder group characterised these vulnerabilities based on their own experiences and observations.

Clinicians

Across the interviews, clinicians referred to several themes, including high-risk clinical situations, misinterpretation and overreliance, pseudo-empathy and emotional vulnerability lack of professional guidance for safe evaluation. They described particular risks for patients with acute suicidality, psychosis, personality disorders, trend diagnoses, or hypochondriacal tendencies, with C01 noting vulnerabilities in "acute suicidality" and situations where AI responses unintentionally reinforced delusional beliefs. Clinicians reported that users may rely too heavily on AI outputs without questioning them, and C03 explained that patients can accept recommendations as certain because the system "formulates its recommendations as if it were completely certain." Interviewees also described emotional risks such as false security, dependency, or substituting AI for interpersonal support, with C02 highlighting that people may use the tools as "pseudo-therapy" or misinterpret responses under stress. C04 described vulnerabilities in systems that cannot reliably distinguish between stress and danger, noting cases where clearly suicidal messages were interpreted as harmless, and mentioned risks of "pseudo-empathy" that can feel supportive but lack real relational grounding. Clinicians also stated they do not feel sufficiently informed to evaluate such tools, with C01 describing their current role as "none" and C04 noting an information gap due to the fast-moving market and absence of a shared clinical framework.

Industry Experts:

Across the interviews, industry experts referred to several themes, including technical weak points, distress-driven user vulnerability, overconfidence and false empathy, and safeguard limitations. They described vulnerabilities appearing during development and deployment, noting that small teams may miss bugs and that incomplete evaluation can allow issues to surface only once systems are in use. IE02 highlighted the role of data quality and underweighted edge cases in producing unreliable outputs and mentioned confusion caused by unclear data-sharing rules. Experts also pointed to user-side risks, with IE03 stating that "distress changes everything" because users become more suggestible and safety classifiers may miss brief or fragmented crisis signals. Interviewees described risks such as hallucinations, overconfident responses, and "false empathy," where emotionally attuned language may invite misplaced trust. IE04 mentioned expectation management as a recurring weak point and noted that users may project more competence onto systems than they possess. They also described remaining vulnerabilities despite established safeguards such as QMS processes and clinical evidence requirements, including emotional escalation, blurred boundaries between support and therapy, and interactions that could be interpreted as therapeutic advice.

Regulators:

Across the interviews, regulators referred to several themes, including psychological harm limits, medical–non-medical grey zones, legal abstraction and pace, and missing interdisciplinary input. They described difficulties in regulating emotional or relational harm, with R01 noting the need to prevent systems from "exploit[ing] people in a vulnerable position or aggravate existing power imbalances." Interviewees stated that frameworks often focus on technical accuracy while issues such as dependency, misplaced trust, or reinforcement of harmful beliefs are harder to capture, as R02 highlighted. Regulators also described unclear boundaries between well-being support and medical functions, which R01 said complicates decisions about referral or prohibition. They noted that legal requirements can be broad or develop slowly, with R03 describing obligations like explainability and oversight as "quite abstract," and R02 mentioning a "pacing problem" due to limited real-world evidence. Interviewees further pointed to insufficient integration of psychological and clinical expertise into safety design, with R01 stating that such input is not consistently included in risk assessments.

Patients:

Across the interviews, patients referred to several themes, including generic or mistuned responses, misreading of emotional intensity, overanalyses or diagnostic drift, and reinforcing rumination. They described moments where answers felt too standard or impersonal, with P01 noting that suggestions such as "do sports" or "write in a journal" felt like a checklist and could come across as dismissive when they were "very desperate." Interviewees also reported that the system often failed to recognise the severity of their emotions, with P01 stating that it could not judge the difference between feeling bad and being "on the verge of doing something impulsive and harmful." Patients described instances where responses leaned too far into analysis, as when P03 wrote "I feel empty today" and received a detailed explanation of depressive disorders, which they had not been seeking. P03 also mentioned becoming more overwhelmed when the AI provided "too many possible interpretations" during moments of distress. Additional vulnerabilities included moments where reassurance was insufficiently tailored or where the system mirrored worry instead of interrupting it, as P03 described when it "doesn't always recognize when I'm stuck in rumination."

4.3 Cross-Layer Coordination (Interdependencies):

The interviews highlighted how different stakeholders described interactions and handovers between regulatory, developer, clinical, and user layers. Participants reported various forms of collaboration, communication, and role coordination, as well as areas where these interactions were experienced as limited or unclear. The following paragraphs summarise how each stakeholder group characterised these interdependencies.

Clinicians

Across the interviews, clinicians referred to several themes, including limited collaboration, unclear responsibilities, and need for structured clinical involvement. They described collaboration between therapists and developers as rare, with C04 explaining that many commercial tools are built with "very little clinical input" and that psychological concepts are sometimes "taken from textbooks and wrapped in a nice interface." Interviewees noted misunderstandings among platforms, clinicians, and users when limitations are not communicated clearly, and C02 stated that this can create "gaps that can become problematic, especially in vulnerable phases." Clinicians also described unclear accountability when incidents occur, with C04 asking whether responsibility lies with "the app provider, the model provider, the app store, or the therapist." Interviewees emphasised the importance of integrating

clinical expertise, with C01 mentioning mandatory involvement of psychotherapists or psychiatrists and C03 stating that a qualified professional should be able to review assessments and intervene if needed. They also mentioned ideas for improving coordination, such as early co-design practices, supervised testing environments, and standardised reporting channels to feed concerns back into the system.

Industry Experts:

Across the interviews, industry experts referred to several themes, including early alignment, communication gaps, responsibility diffusion, and role definition. They described structured interaction with regulators, clinicians, and payers through evidence submissions, advisory boards, and post-market surveillance, with IE01 noting that, the company he works for builds trust through "scientific rigor" and strong clinical evidence. Interviewees reported limited direct exchange in some settings, with IE02 explaining that developers were often "left to follow the project specifications on their own" and sometimes misunderstood data sharing rules. Experts also mentioned accountability gaps when multiple actors are involved, with IE03 describing "responsibility diffusion" between model and application providers and noting that misaligned prompts can leave "users fall[ing] into the gap." They added that coordination is shaped by differing incentives, as IE04 stated that engineering, clinical, and regulatory teams optimise for different priorities and that product teams must "hold the centre." Interviewees also highlighted the need for shared terminology, with IE03 stating that stakeholders still lack common language for emotional risk, escalation, and unsafe advice, which complicates alignment across layers.

Regulators:

Across the interviews, regulators referred to several themes, including fragmented collaboration, clinical–technical disconnect, accountability gaps, transparency and resource limits, and jurisdiction differences. They described collaboration as inconsistent, with R01 noting that despite consultation processes and expert groups, there is "a lack of institutionalised collaboration" and that legal expectations, technical implementation, and clinical realities "do not always meet in a coherent way." Interviewees highlighted limited and unsystematic involvement of clinicians or psychologists in development, and R02 stated that there is "no formal mechanism guaranteeing clinical involvement" for high-risk conversational systems. Regulators also pointed to unclear responsibilities when multiple actors are involved, with R02 noting that when model and application providers differ, "accountability becomes unclear." Interviewees described gaps in transparency regarding training data, risk assessments, and

interpretations of requirements, and R02 referred to a “structural knowledge gap” linked to limited regulatory resources. They additionally mentioned challenges arising from differing national and international requirements, which R01 and R03 associated with inconsistent safeguards across jurisdictions.

Patients:

Across the interviews, patients referred to several areas, including Partial Trust, Using AI Before Humans, and Potential Delays in Seeking Help. They described trusting AI for information and structure but not for assessing risk or understanding their situation, with P01 noting that real relationships offer something “the AI lacks.” Interviewees also mentioned using AI as a first step before contacting professionals, as P02 explained that they use it “for a first impression” but not as a replacement for diagnosis. Patients described relying more heavily on the AI during difficult phases, and P01 stated that they sometimes isolated themselves because it felt “easier to confide in the AI.” They also discussed reassurance seeking, with P02 asking the same questions repeatedly “on the same day.” Interviewees mentioned that systems could contribute to delayed help seeking, and P03 said that people may rely on the chatbot as a “safe middle ground” and convince themselves they are managing even when professional care would be more appropriate.

4.4 Safeguards & Governance

The interviews highlighted a range of safeguards and governance measures that participants considered important for the safe use of AI-based mental health tools. Across clinicians, industry experts, regulators, and patients, interviewees described mechanisms, limits, and oversight structures they viewed as necessary to reduce risks and support safe interactions. The following paragraphs summarise how each stakeholder group characterised these safeguards.

Clinicians:

Across the interviews, clinicians referred to several themes, including supplement not replacement, clear limits and crisis protocols, and professional oversight. They stated that AI should support but never replace therapy, with C01 and C04 describing its role in psychoeducation, journaling, skill practice, or as a bridge during waiting periods, while emphasising that severe crises or complex conditions require human-led care. Interviewees stressed the need for explicit limits and reliable crisis detection, and C02 highlighted safeguards such as clear labelling, transparency about capabilities, warnings, and keyword-based

escalation pathways. Clinicians also emphasised data protection, clinical studies, certification, and continuous monitoring, and C03 noted that qualified professionals should review system accuracy and remain involved because mental health lacks clear yes or no diagnostic categories. C04 additionally described staged access models, explicit "exit signs," and monitoring for overuse or isolation as ways to prevent pseudo-intimacy and proposed a defined "handover point" where systems guide users to human support once symptoms, repeated crisis cues, or stagnation indicate the need for professional care.

Industry Experts:

Across the interviews, industry experts referred to several themes, including clinical-grade constraints, evidence and oversight, and risk-sensitive technical controls. IE01 described the need for guideline-constrained models that prevent hallucinations and require clinical-grade evidence comparable to earlier DTx systems, adding that prescribers and patients need better psychoeducation. Interviewees emphasised strong development oversight, with IE01 highlighting ethical reviews and evidence-driven decisions, and IE04 noting that nothing should be released without "clinical sign-off" and clear regulatory expectations. Experts also pointed to practical safeguards such as encryption, employee training, and transparent data-handling briefings, which IE02 said help users make informed choices about what they share. They additionally described technical measures for crisis scenarios, with IE03 stating that systems should "detect, de-escalate, hand off" and avoid improvisational empathy, and proposed hybrid pathways where a separate module manages escalation. Interviewees discussed emerging safety approaches such as risk-sensitive decoding, behavioural constraints, continuous monitoring, and audit trails, and IE04 cautioned that personalised companions and autonomous agents create attachment dynamics and compounding risk.

Regulators:

Across the interviews, regulators referred to several themes, including liability and transparency, independent oversight, and psychological safety standards. They described the need for clear liability rules and stronger incentives for firms to prioritise user safety, with R01 stating that psychological and emotional harms can be "just as serious as physical harms." Interviewees emphasised transparency obligations such as documenting system limitations, data risks, and safeguards for vulnerable users, and R01 noted that risk assessments should address "vulnerability, power imbalances and resilience." Regulators also highlighted independent audits and continuous monitoring, with R02 pointing to the importance of tracking

how systems behave with real users and R03 describing interdisciplinary audits that examine bias, mental privacy and autonomy. They additionally discussed embedding psychological safety into regulation, and R02 said it should be "explicit rather than implied," while R03 called for clearer rules for mental-health contexts and enforceable rights to human review. Interviewees also mentioned emerging high-risk use cases such as emotion-profiling tools, adaptive chatbots and behavioural monitoring systems, which R01 described as sitting at the intersection of "power imbalances, dependency and limited resilience."

Patients:

Across the interviews, patients referred to several themes, including crisis differentiation, clear boundaries, and harm-avoidance rules. They described the need for systems to distinguish between information seeking and acute crisis, with P01 stating that when language suggests despair or self-harm, the AI should "clearly state that it cannot replace emergency help" and immediately refer to real support. Interviewees emphasised explicit limits, such as repeated reminders that the AI is not a professional, and P03 said that messages like "I'm not a therapist, but here are some basic strategies" would set needed boundaries. Patients also highlighted safeguards to prevent harmful outputs, noting that chatbots should never give instructions for self-harm, minimise feelings, dramatise symptoms, or provide diagnostic labels, and P02 cautioned that dramatic or heavy phrasing can intensify anxiety. They additionally mentioned features such as small actionable steps, automatic pointers to crisis services, protection against manipulation, and transparency about data handling, with P03 saying they want to know "what the AI does with what I share."

Topic	Clinicians	Industry Experts	Regulators	Patients
Responsibilities & Controls	• Safety responsibility is shared, but users should not carry the main burden. • Clear limits and transparency are essential.	• Roles should be clearly separated across regulators, developers, and clinicians. • Developer responsibility is limited after deployment.	• Regulators set minimum safety standards; developers must operationalise them. • User responsibility is inherently limited.	• Users cannot self-regulate reliably when distressed. • Safety should be ensured by developers and regulators.
Vulnerabilities	• High-risk states (e.g. suicidality) amplify over-reliance and misinterpretation. • Pseudo-empathy creates false reassurance.	• Technical gaps and data limits expose users in distress. • Authoritative tone can invite misplaced trust.	• Psychological and relational harms are hard to capture legally. • Medical vs non-medical boundaries remain unclear.	• Emotional intensity is often misread. • Reassurance loops and rumination may be reinforced.
Cross-Layer Coordination	• Clinical input in development is limited. • Accountability across actors is unclear.	• Coordination gaps exist between model, app, and regulator levels. • Incentives and language are misaligned.	• Legal, technical, and clinical layers remain fragmented. • Multi-actor accountability is unclear.	• AI is often used before contacting humans. • Reliance on AI may delay help-seeking.
Safeguards & Governance	• AI should supplement, not replace, therapy. • Clear crisis escalation is required.	• Guardrails, constrained generation, and monitoring are essential. • Evidence and audits should precede release.	• Stronger liability and psychological safety standards are needed. • Mental-health uses warrant high-risk treatment.	• Crisis situations must trigger referral to human help. • Clear boundaries and harm-avoidance are expected.

Table 2: Summary Results Table

5. Discussion

This chapter interprets the empirical findings in relation to the study's overarching research question and subsidiary questions, using the layered patient-safety perspective developed in the Background. Structured around SQ1–SQ4, the chapter synthesises convergences, tensions, and mechanisms across regulatory, developer, clinical, and patient layers, and uses these insights to refine the layered patient-safety framework by clarifying where safety is lost, how vulnerabilities align at interfaces, and under which conditions layered safeguards succeed or fail.

5.1 Safety Responsibilities as a Distributed but Tensioned System (SQ1)

Across stakeholder groups, a strong consensus emerges that patient safety in AI-driven mental-health tools cannot be ensured by any single actor. Instead, responsibility is understood as inherently distributed across regulatory, developer, clinical, and patient layers. Crucially, however, the findings indicate that safety risk does not arise from the absence of responsibility, but from misalignment in how responsibility is interpreted, enacted, and handed over between layers. This pattern is consistent with recent work on layered accountability in AI-based mental-health interventions, which highlights that diffuse responsibility becomes a safety risk when roles are not operationally aligned across system layers (Lei, 2025).

At a system level, all stakeholders reject models that place primary responsibility on patients. Vulnerable users are consistently framed as recipients of protection rather than bearers of safety obligations. This shared position reflects a foundational patient-safety assumption: during psychological distress, individuals cannot be expected to critically assess system reliability or manage risk autonomously. Similar concerns are raised in system-level assessments of AI readiness in healthcare, which emphasise that patient-facing AI systems must not rely on individual user judgment as a primary safety mechanism (WHO, 2025). The convergence here is notable, as it establishes user responsibility not as a safeguard but as a potential failure point when relied upon excessively.

Where divergence emerges is in how upstream responsibility is conceptualised and operationalised. Regulators, developers, and clinicians each operate according to distinct logics of safety. Regulatory responsibility is framed in abstract and rights-based terms, focused on classification, compliance, and baseline protections. Developer responsibility is operational and design-oriented, emphasising lifecycle control, guardrails, and risk mitigation through technical and interface choices. Clinical responsibility, by contrast, is situational and judgment-based,

centred on contextual interpretation, escalation, and therapeutic boundaries. While each logic is internally coherent, they do not naturally align.

This misalignment creates a structural tension: each layer assumes that safety failures will be intercepted elsewhere. Regulators define guardrails but rely on developers to interpret and implement them correctly. Developers embed safeguards but assume clinical oversight will manage contextual risk. Clinicians expect tools made available to patients to already be safe by design. These assumptions are individually reasonable, yet collectively hazardous. Interpreted through a layered safety lens, responsibility becomes diffuse rather than cumulative, allowing gaps to emerge where no actor holds end-to-end accountability. Similar dynamics of responsibility diffusion have been observed in analyses of AI governance in mental health, where accountability is formally distributed but practically fragmented (Lei, 2025).

A particularly salient contradiction arises around boundary-setting. Developers and regulators emphasise explicit limits and disclaimers to prevent overreach, while clinicians focus on situational thresholds that depend on patient state rather than system rules. This difference reflects a deeper tension between standardisation and discretion. When technical boundaries fail to map onto clinical realities, or when clinicians encounter tools without clear design intent, safety responsibilities become blurred rather than reinforced.

Importantly, these tensions do not represent disagreement over the importance of safety, but rather conflicting assumptions about where safety should be actively maintained. The findings suggest that safety deteriorates when responsibility is treated as transferable rather than relational. In other words, when layers delegate rather than coordinate, responsibility gaps function as latent vulnerabilities within the system.

From a patient-safety perspective, this analysis supports the hypothesis that distributed responsibility is protective only when roles are explicitly synchronised. Safety emerges not from delegation, but from alignment: regulators setting enforceable baselines, developers translating them transparently into system behaviour, clinicians retaining authority to intervene, and patients being guided rather than burdened. Where these roles are misaligned, even well-intentioned safeguards can fail.

In sum, SQ1 demonstrates that patient safety in conversational mental-health AI depends less on defining who is responsible than on ensuring that responsibility is continuously coordinated across layers. Without such coordination, responsibility itself becomes a source of risk rather than protection. The findings largely support Hypothesis 1, while refining it by showing that patient safety risks arise less from disagreement over responsibility than from implicit delegation and insufficient coordination of responsibilities across layers.

5.2 Layered Vulnerabilities and the Alignment of Risk (SQ2)

The findings indicate that safety vulnerabilities in AI-driven mental-health tools arise not from isolated failures within individual layers, but from the alignment of otherwise manageable weaknesses across regulatory, technical, clinical, and user contexts. Interpreted through the Swiss Cheese Model, these vulnerabilities function as latent weaknesses that become consequential when they coincide, particularly under conditions of psychological distress. This systemic pattern aligns with prior reviews showing that harms in AI-based mental-health interventions typically emerge from interactions between model limitations, deployment conditions, and user vulnerability rather than from single technical errors (Ali et al., 2025; Löchner et al., 2025).

A recurring vulnerability concerns risk recognition and differentiation. Across layers, there is limited capacity to reliably distinguish between low-risk emotional support and high-risk clinical situations. Conversational systems struggle to detect severity, clinicians lack formal criteria for evaluating AI-mediated interactions, and regulatory frameworks insufficiently capture psychological escalation. Similar limitations have been identified in scoping and systematic reviews, which note that AI-driven mental-health tools often fail to reliably signal when clinical or emergency intervention is required (Ni & Jia, 2025; Guo et al., 2024).

A second vulnerability emerges around pseudo-empathy and authority ambiguity. Conversational systems frequently employ affirming language that can be interpreted as therapeutic validation, increasing trust without corresponding clinical judgment. This mirrors concerns raised in critical literature describing how simulated empathy may reinforce maladaptive beliefs or emotional dependence in vulnerable users (Lei, 2025; Denecke et al., 2025).

A third vulnerability relates to overreliance driven by accessibility and continuity. While constant availability lowers barriers to support, it also enables prolonged engagement without enforced transition to human care. Prior studies similarly highlight delayed help-seeking as an unintended consequence when escalation mechanisms remain implicit or user-driven (Denecke et al., 2025; Löchner et al., 2025).

Finally, evaluation and oversight gaps further amplify risk. Clinicians report limited guidance for assessing AI tools, regulators acknowledge difficulty monitoring real-world conversational behaviour, and developers lack visibility into downstream use. This reflects broader observations that governance and evaluation structures have not yet adapted to the adaptive and relational nature of generative AI systems in mental health (Ali et al., 2025; WHO,2025).

Taken together, SQ2 demonstrates that patient-safety risks in conversational mental-health AI are systemic rather than accidental. Vulnerabilities intensify where responsibility, authority, and perception intersect, particularly for users least able to compensate for system weaknesses. In line with Hypothesis 2, the findings show that risk emerges primarily through the alignment of vulnerabilities across layers rather than through isolated failures within any single domain.

5.3 Cross-Layer Interdependencies as Systemic Safety Determinants (SQ3)

Interpreting the findings through a layered patient-safety lens reveals that risks in conversational mental-health AI arise less from failures within individual layers than from how those layers interact. Across stakeholder perspectives, interdependencies emerge as systemic safety determinants: when coordination between regulatory, developer, clinical, and patient layers is explicit and structured, risks are buffered; when it is informal, delayed, or ambiguously owned, vulnerabilities align and amplify. This pattern supports the hypothesis that safety outcomes are shaped primarily at interfaces rather than within isolated domains. Similar interface-driven risk dynamics have been highlighted in prior analyses of AI-based mental-health interventions, which emphasise fragmented accountability and weak cross-sector coordination as central safety challenges (Balcombe, 2023; WHO, 2025; Denecke et al., 2025).

A first point of convergence across stakeholders concerns the timing and integration of expertise. Clinicians, developers, and regulators implicitly agree that safety is compromised when clinical insight is introduced only after core technical decisions have been made. However, they diverge in how they interpret this misalignment. Clinicians view late involvement as a loss of therapeutic calibration, developers frame it because of innovation cycles, and regulators treat it as a governance gap rather than a design flaw. These differing interpretations obscure accountability: although all actors recognise the risk, none clearly owns responsibility for closing it. At the system level, safeguards may exist in principle yet fail to interlock because they are developed sequentially rather than collaboratively, resulting in interface failures rather than isolated breakdowns.

A second shared pattern concerns responsibility diffusion in modular AI ecosystems. Stakeholders consistently describe conversational AI systems as distributed across multiple entities, including model providers, application developers, and deployers. While modularity enables scalability and innovation, it simultaneously fragments safety ownership. Developers assume clinicians will contextualise use, clinicians assume unsafe tools would not be deployed, and regulators assume compliance will be interpreted conservatively. These assumptions are individually reasonable yet collectively dangerous, as they create conditions in which no layer

intervenes decisively when risk escalates. This observation aligns with recent reviews of large language models in mental health, which identify modular architectures and multi-actor deployment as recurring sources of accountability gaps and safety blind spots (Guo et al., 2024; Ali et al., 2025).

A third interdependency emerges around interpretive alignment, particularly regarding how AI systems are perceived by users. Stakeholders agree that conversational agents are often interpreted as quasi-therapeutic, even when they are not intended or regulated as such. Responsibility for this misperception is contested: developers emphasise disclaimers and functional limits, clinicians focus on the risk of role confusion, and regulators highlight classification gaps. The result is a semantic misalignment in which different layers operate with incompatible understandings of "support," "therapy," and "appropriate intervention." Prior ethical analyses similarly warn that simulated empathy and conversational framing can blur authority boundaries, fostering misplaced trust and delayed escalation in vulnerable users (Denecke et al., 2025; Lei, 2025).

Patients experience these coordination failures most acutely through opaque transitions to human care. While AI tools are often used as an accessible first step, users frequently lack clarity about when and how responsibility shifts from system to professional. Although stakeholders broadly agree that patients should not manage escalation independently during distress, escalation pathways are often weakly enforced or left implicit. This gap between normative agreement and operational reality allows vulnerability to persist unmitigated, echoing findings from scoping reviews that identify the absence of enforceable handover mechanisms as a central weakness in current AI-based mental-health interventions (Ni & Jia, 2025)

Taken together, SQ3 demonstrates that interdependencies are not peripheral to patient safety but constitutive of it. Failures arise not because individual layers neglect their roles, but because safety-relevant assumptions are made in isolation and left uncoordinated. Interfaces become failure points when responsibility is implicit rather than assigned, meanings are unshared, and escalation relies on expectation rather than enforcement. In Swiss Cheese terms, hazards pass through not because layers are absent, but because their holes align through weak coordination. These findings largely support Hypothesis 3, while refining it by showing that safety failures at interfaces are driven less by overt conflict than by uncoordinated assumptions and interpretive misalignment across layers.

5.4 Safeguards and Governance as Layered Safety Mechanisms (SQ4)

Interpreted across stakeholder perspectives, safeguards and governance mechanisms are understood not as isolated technical controls, but as systemic arrangements that constrain behaviour, clarify authority, and enable timely human intervention. A central analytical pattern is that safeguards are considered effective only when they operate across layers and, crucially, at their interfaces. This aligns with recent literature arguing that psychological safety in AI-based mental-health interventions depends on layered governance structures rather than single-point controls (Balcombe, 2023; Lei, 2025).

Across all groups, there is strong convergence on the idea that boundaries constitute a primary safety mechanism. Clinicians, developers, regulators, and patients alike emphasise the importance of preventing role confusion between AI support and psychotherapy. However, stakeholders diverge in how such boundaries should be enforced. Clinicians frame boundaries as professional limits requiring explicit stopping rules and handovers, while industry experts focus on design constraints and behavioural limits embedded in system architecture. Regulators emphasise boundary enforcement through accountability and classification, and patients focus on whether limits are visible and felt during moments of vulnerability. This mirrors findings from prior reviews showing that boundary-setting fails when it is implemented at only one layer rather than coordinated across technical, clinical, and regulatory domains (Denecke et al., 2025; Ali et al., 2025).

A second shared pattern concerns escalation and stopping mechanisms. Stakeholders broadly agree that continuous conversational availability without enforced escalation increases risk, particularly for users experiencing prolonged or intensifying distress. Clinicians interpret this as a failure of therapeutic containment, industry experts as a governance lag expanding system capabilities, and regulators as a gap in enforceable standards. Patients describe uncertainty about when the system can no longer help and when human support is required. This finding is consistent with scoping reviews that identify delayed escalation and weak handover pathways as recurrent safety failures in AI-driven mental-health tools (Ni & Jia, 2025; Guo et al., 2024).

At the same time, stakeholders reveal tensions around how much constraint is appropriate. Industry experts and regulators express concern that overly rigid safeguards could undermine accessibility, innovation, or early intervention potential. Clinicians and patients, by contrast, prioritise protection over flexibility in high-risk scenarios. This reflects a broader governance dilemma identified in the literature: safeguards must balance proportionality and protection, scaling with both system capability and user vulnerability rather than applying

uniform rules (Löchner et al., 2025; WHO, 2025). Static safeguards are therefore insufficient for dynamic, adaptive conversational systems.

A further divergence emerges around where responsibility for safeguarding should ultimately reside. Regulators conceptualise safeguards primarily as accountability structures, emphasising liability, documentation, and continuous monitoring. Industry experts focus on embedding safety upstream through constrained generation and capability limits, while clinicians stress the irreplaceability of professional judgment in ambiguous psychological contexts. Patients locate safety in interactional cues that signal care, limits, and responsiveness. Rather than contradicting one another, these perspectives illustrate that safeguards operate at legal, technical, clinical, and experiential levels simultaneously. Prior ethical analyses similarly argue that safety failures occur when governance mechanisms privilege one level while neglecting others (Lei, 2025).

Synthesising these insights, SQ4 demonstrates that safeguards are most effective when treated as interface mechanisms rather than isolated controls. They work by preventing vulnerabilities from aligning across layers, not by eliminating risk within a single domain. In Swiss Cheese terms, governance does not remove holes; it staggers them. Effective safety architectures therefore combine constraint-based design, enforceable accountability, clinical oversight, and visible patient-facing limits into a coherent, layered system.

This interpretation reinforces the broader argument of the thesis: safeguarding conversational AI in mental health requires governance structures that are layered, adaptive, and explicitly oriented toward psychological safety. When safeguards are fragmented or implicit, they fail to interrupt risk trajectories. When aligned across layers and interfaces, they enable innovation while preserving protection for vulnerable users. These findings partially support Hypothesis 4, while also highlighting persistent tensions between protection, accessibility, and innovation that complicate consistent implementation.

Taken together, the discussion of SQ1–SQ4 provides a structured interpretation of how patient safety in AI-driven conversational mental-health tools is shaped across regulatory, technical, clinical, and user layers. Each sub-question illuminates a distinct dimension of the safety problem, responsibility allocation, vulnerability formation, cross-layer coordination, and safeguarding mechanisms, while also clarifying how these dimensions interact at the interfaces between layers. In doing so, the discussion refines the layered patient-safety framework by specifying where safety-relevant misalignments emerge and under which conditions vulnerabilities align to create risk.

At the same time, the empirical findings also point to a set of broader, closely related issues that extend beyond patient safety in a narrow sense. These include the role of psychological vulnerability, the ethical implications of emotionally responsive AI, questions of individual and institutional responsibility, and the wider societal consequences of deploying emotionally impactful technologies in situations of distress.

5.5 Broader Considerations Beyond Patient Safety

Beyond patient safety, the findings highlight psychological vulnerability as a central contextual condition shaping how conversational AI systems are interpreted and relied upon. Vulnerability emerges not as a fixed individual characteristic, but as a fluctuating state influenced by distress, isolation, and uncertainty. Emotionally responsive systems interact directly with these states, affecting users' judgment, trust calibration, and help-seeking behaviour. As a result, identical system behaviour can produce very different outcomes depending on users' emotional capacity at the moment of interaction.

In addition, the study contributes to debates on the ethics of emotionally responsive AI. By simulating empathy and care, conversational systems blur boundaries between information provision, emotional support, and therapeutic interaction. While such design choices may increase engagement and disclosure, they also risk creating false impressions of understanding or authority. Ethical concerns thus arise not only from technical failure, but from the tension between scalable emotional responsiveness and the absence of genuine relational accountability.

Finally, the findings point to broader societal implications. As conversational AI becomes embedded in everyday coping practices, it may reshape norms around help-seeking and responsibility for mental well-being. While such tools may partially address access gaps, they also risk normalising individualised, technology-mediated coping in contexts of constrained healthcare capacity, raising questions about whether responsibility for care is being redistributed rather than resolved.

Together, these considerations show that safeguarding users of conversational AI requires attention not only to patient safety mechanisms, but also to vulnerability, ethics, and societal context. Addressing these dimensions alongside safety strengthens the interpretive framework of this thesis and situates patient safety within the wider ethical and social landscape of emotionally responsive AI in mental-health settings.

5.6 Limitations

This study has several limitations that should be acknowledged. First, the layered patient-safety framework proposed in this thesis was developed through conceptual integration and qualitative interpretation rather than formal empirical validation. This was not critical to the study's aims, as the objective was to refine existing safety concepts for an emerging domain rather than to test a predictive model. Second, the study relies on a purposive qualitative sample, and access to certain stakeholder groups was constrained by availability and the sensitivity of the topic. While this limited sample size and balance, it was appropriate given the exploratory nature of the research and the difficulty of recruiting expert and vulnerable-user perspectives. Finally, the regulatory and technological landscape for conversational mental-health AI is evolving rapidly, which may affect the generalisability of specific governance observations. However, the layered framework developed in this thesis is intended to be adaptive rather than static and can be refined as empirical evidence and regulatory practice mature.

6. Conclusion

This thesis set out to examine how patient safety in AI-driven conversational mental-health tools can be conceptualised and safeguarded through a layered framework. Drawing on qualitative insights from clinicians, industry experts, regulators, and patients, the study demonstrates that patient safety in this context cannot be reduced to technical accuracy, regulatory compliance, or individual user responsibility. Instead, safety emerges as a system-level property of a layered socio-technical arrangement, shaped by how regulatory, technical, clinical, and user layers interact, align, and compensate for one another.

In response to the research question of how patient safety in AI-driven mental-health tools, particularly conversational agents, can be conceptualised and safeguarded through a layered framework, this study shows that safety is best understood as an emergent property of the system as a whole. Drawing on the Swiss Cheese Model, the findings demonstrate that patient harm does not typically arise from isolated failures within individual layers, but from the alignment of vulnerabilities across regulatory, technical, clinical, and user domains, particularly in moments of psychological distress. Many of the weaknesses identified by participants, such as limitations in crisis recognition, over-reliance on empathic language, delayed escalation, or fragmented accountability, were described as understandable or even unavoidable within their respective contexts. However, when these vulnerabilities coincide without effective interruption at the interfaces between layers, the risk of harm escalates. Safeguarding patient safety therefore depends less on eliminating individual weaknesses than on preventing their convergence through coordinated responsibility, clear escalation pathways, and aligned governance across layers.

Several implications follow from this system-level understanding of patient safety. First, safety cannot be ensured through single-point interventions. Improvements limited to one layer, such as more accurate algorithms, clearer disclaimers, or stricter regulatory oversight, remain insufficient if adjacent layers remain misaligned. Effective safety architectures must therefore prioritise coordination mechanisms alongside individual safeguards. Second, psychological harm must be recognised as a central safety outcome rather than a peripheral concern. Risks such as emotional dependency, misplaced trust, delayed help-seeking, and pseudo-empathy emerged as core modes of harm in conversational mental-health AI, particularly for vulnerable users, and require explicit consideration within safety frameworks. Third, the findings challenge the reliance on user responsibility as a primary safety mechanism. Periods of psychological distress are precisely those in which users are least able to interpret system limitations, calibrate trust, or initiate escalation independently. Safety designs must therefore

anticipate impaired judgement and proactively guide users toward appropriate forms of support. More broadly, the analysis highlights the importance of proportional and adaptive governance. Safeguards must scale with both system capability and user vulnerability, balancing accessibility and early support against the risk of harm, rather than relying on either permissive or overly restrictive approaches in isolation.

Several limitations of this study should be acknowledged. The findings are based on a qualitative sample and reflect stakeholder perceptions rather than measured safety outcomes. While this approach is appropriate for exploring an emerging and under-theorised domain, it does not allow for causal inference or quantification of risk. In addition, regulatory environments and technological capabilities are evolving rapidly, meaning that the analytical framework developed here should be understood as adaptive rather than static and as reflecting current practices rather than definitive safety standards.

Despite these limitations, this thesis contributes to the existing research by offering a holistic and system-oriented conceptualisation of patient safety in AI-driven conversational mental-health tools. By integrating regulatory, technical, clinical, and user perspectives within a single layered framework, the study brings together insights that have often been treated in isolation. In doing so, it demonstrates that improving coordination and alignment between layers is more critical for patient safety than strengthening any single layer in isolation.

Future research could build on these findings by examining layered safety mechanisms through longitudinal studies, experimental designs, or real-world deployment data. Further work is also needed to operationalise interface governance, particularly with regard to escalation thresholds, responsibility transfer, and human–AI collaboration in moments of acute distress. Comparative research across regulatory contexts and system designs may additionally support the development of context-sensitive and proportionate safety standards for conversational mental-health AI.

Bibliography

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics, 132, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

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, 20552076251395548. https://doi.org/10.1177/20552076251395548

Baek, G., Cha, C., & Han, J. (2025). AI chatbots for psychological health for health professionals: Scoping review. JMIR Human Factors, 12, e67682. https://doi.org/10.2196/67682

Balan, R., Dobrean, A., & Poetar, C. R. (2024). Use of automated conversational agents in improving young population mental health: a scoping review. NPJ digital medicine, *7*(1), 75. https://doi.org/10.1038/s41746-024-01072-1

Balcombe, L. (2023). AI Chatbots in Digital Mental Health. Informatics, 10(4), 82. https://doi.org/10.3390/informatics10040082

Beauchamp, T., & Childress, J. (2019). Principles of biomedical ethics: Marking its fortieth anniversary. The American Journal of Bioethics, *19*(11), 9–12. https://doi.org/10.1080/15265161.2019.1665402

Bengio, Y., et al. (2024). Managing extreme AI risks amid rapid progress. Science, 384, 842–845. https://doi.org/10.1126/science.adn0117

Blease, C., & Torous, J. (2023). ChatGPT and mental healthcare: Balancing benefits with risks of harms. BMJ Mental Health, 26(1), e300884. https://doi.org/10.1136/bmjment-2023-300884

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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, Article 100017. https://doi.org/10.1016/j.ssmmh.2021.100017

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

Chakraborty, C., Pal, S., Bhattacharya, M., Dash, S., & Lee, S. S. (2023). Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Frontiers in Artificial Intelligence, 6, 1237704. https://doi.org/10.3389/frai.2023.1237704

Choudhury, A., & Asan, O. (2020). Role of artificial intelligence in patient safety outcomes: Systematic literature review. JMIR Medical Informatics, 8(7), e18599. https://doi.org/10.2196/18599

Chu, M., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025). *Illusions of intimacy: Emotional attachment and emerging psychological risks in human–AI relationships*. arXiv. https://doi.org/10.48550/arxiv.2505.11649

CNN Business. (2025, March 17). Parents of 16-year-old Adam Raine sue OpenAI, claiming ChatGPT advised on his suicide. CNN Business. https://edition.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit

Cummings, M. (2024). A Taxonomy for AI Hazard Analysis. Journal of Cognitive Engineering and Decision Making, 18, 327 - 332. https://doi.org/10.1177/15553434231224096.

Dehbozorgi, R., Zangeneh, S., Khooshab, E., et al. (2025). The application of artificial intelligence in the field of mental health: A systematic review. BMC Psychiatry, 25, 132. https://doi.org/10.1186/s12888-025-06483-2

Denecke, K., May, R., Gabarron, E., & Lopez-Campos, G. H. (2023). Assessing the potential risks of digital therapeutics (DTX): The DTX risk assessment canvas. Journal of Personalized Medicine, 13(10), 1523. https://doi.org/10.3390/jpm13101523

Denecke, K., Lopez-Campos, G., & Gabarron, E. (2025). Hidden in plain sight: The harmful side of AI-based mental health interventions. Studies in Health Technology and Informatics, 327, 248–252. https://doi.org/10.3233/SHTI250322

Dohnányi, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2025). Technological folie à deux: Feedback loops between AI chatbots and mental illness. arXiv preprint arXiv:2507.19218. https://doi.org/10.48550/arXiv.2507.19218

European Parliament & Council of the European Union. (2024, July 12). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (Artificial Intelligence Act)*. *Official Journal of the European Union, L 1689*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Floridi, L. (2024, May 24). The ethics of artificial intelligence: Exacerbated problems, renewed problems, unprecedented problems - Introduction to the special issue of the American Philosophical Quarterly dedicated to the ethics of AI. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4801799

Gabriel, S., Puri, I., Xu, X., Malgaroli, M., & Ghassemi, M. (2024). Can AI relate: Testing large language model response for mental health support. arXiv preprint arXiv:2405.12021. https://doi.org/10.48550/arXiv.2405.12021

Gamble, A. (2020). Artificial intelligence and mobile apps for mental healthcare: A social informatics perspective. Aslib Journal of Information Management, Advance online publication. https://doi.org/10.1108/AJIM-11-2019-0316

Goktas, P., & Grzybowski, A. (2025). *Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI*. Journal of Clinical Medicine, 14, 1605. https://doi.org/10.3390/jcm14051605

Golden, A., & Aboujaoude, E. (2024a). The framework for AI tool assessment in mental health (FAITA – Mental Health): A scale for evaluating AI-powered mental health tools. World Psychiatry, 23(3), 444–445. https://doi.org/10.1002/wps.21248

Golden, A., & Aboujaoude, E. (2024b). 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

Guo, Z., Lai, A., Thygesen, J., Farrington, J., Keen, T., & Li, K. (2024). Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health*, 11. https://doi.org/10.2196/57400.

Hipgrave, L., Goldie, J., Dennis, S., & Coleman, A. (2025). Balancing risks and benefits: Clinicians' perspectives on the use of generative AI chatbots in mental healthcare. Frontiers in Digital Health, 7, Article 1606291. https://doi.org/10.3389/fdgth.2025.1606291

Huang, X., Nor, M., Yusoff, Z., & Liang, Z. (2025). *Psychological harm induced by generative AI and its legal remedies*. Environment and Social Psychology, 10(3). https://doi.org/10.59429/esp.v10i3.3554

Ingle, M. (2025). A review on AI-powered conversational agents for psychotherapy: Techniques, tools, and ethical perspectives – Inner Me. International Journal for Research in Applied Science and Engineering Technology, 13, 5052–5053. https://doi.org/10.22214/ijraset.2025.71307

Iorfino, F., Davenport, T., Ospina-Pinillos, L., Hermens, D., Cross, S., Burns, J., & Hickie, I. (2017). *Using new and emerging technologies to identify and respond to suicidality among help-seeking young people: A cross-sectional study*. Journal of Medical Internet Research, 19, e247. https://doi.org/10.2196/jmir.7897

Jimini Health. (2025, July 8). The new Hippocratic code: An LLM-native safety framework for patient-facing AI in mental health [Blog post]. https://jiminihealth.com/blog/the-new-hippocratic-code-an-llm-native-safety-framework-for-patient-facing-ai-in-mental-health

Kalam, K. T., Rahman, J. M., Islam, M. R., & Dewan, S. M. R. (2024). ChatGPT and mental health: Friends or foes? Health Science Reports, 7(2), e1912. https://doi.org/10.1002/hsr2.1912

Kaywan, P., Ahmed, K., Ibaida, A., Miao, Y., & Gu, B. (2023). Early detection of depression using a conversational AI bot: A non-clinical trial. *PloS one*, *18*(2), e0279743. https://doi.org/10.1371/journal.pone.0279743

Kulkarni, M., Mantere, S., Vaara, E., van den Broek, E., Pachidi, S., Glaser, V. L., Gehman, J., Petriglieri, G., Lindebaum, D., Cameron, L. D., Rahman, H. A., Islam, G., & Greenwood, M. (2024). The Future of Research in an Artificial Intelligence-Driven World. *Journal of Management Inquiry*, *33*(3), 207-229. https://doi.org/10.1177/10564926231219622

Larouzée, J., & Le Coze, J.-C. (2020). *Good and bad reasons: The Swiss cheese model and its critics*. Safety Science, 126, Article 104660. https://doi.org/10.1016/j.ssci.2020.104660

Lee, H. S., Wright, C., Ferranto, J., Buttimer, J., Palmer, C. E., Welchman, A., Mazor, K. M., Fisher, K. A., Smelson, D., O'Connor, L., Fahey, N., & Soni, A. (2025). Artificial intelligence conversational agents in mental health: Patients see potential, but prefer humans in the loop. *Frontiers in psychiatry*, *15*, 1505024. https://doi.org/10.3389/fpsyt.2024.1505024

Lei, X. (2025). The ethical challenges of AI-based mental health interventions: Toward a layered accountability framework. Lecture Notes in Education Psychology and Public Media, 115**.** https://doi.org/10.54254/2753-7064/2025.HT26261

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

Löchner, J., Carlbring, P., Schuller, B., Torous, J., & Sander, L. B. (2025). Digital interventions in mental health: An overview and future perspectives. Internet Interventions, 40, 100824. https://doi.org/10.1016/j.invent.2025.100824

MacNeill, A. L., MacNeill, L., Luke, A., & Doucet, S. (2024). Health professionals' views on the use of conversational agents for health care: Qualitative descriptive study. Journal of Medical Internet Research, 26, e49387. https://doi.org/10.2196/49387

Mansoor, M., Hamide, A., & Tran, T. (2025). Conversational AI in Pediatric Mental Health: A Narrative Review. Children (Basel, Switzerland), 12(3), 359. https://doi.org/10.3390/children12030359

Maurya, R., Montesinos, S., Bogomaz, M., & DeDiego, A. (2024). Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. Counselling and Psychotherapy Research. https://doi.org/10.1002/capr.12759

Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. JMIR mental health, *12*, e60432. https://doi.org/10.2196/60432

Meszaros, J., Minari, J., & Huys, I. (2022). The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. Frontiers in Genetics, 13, 927721. https://doi.org/10.3389/fgene.2022.927721

Migliorini, S. (2024). "More than Words": A Legal Approach to the Risks of Commercial Chatbots Powered by Generative Artificial Intelligence. European Journal of Risk Regulation, 15, 719 - 736. https://doi.org/10.1017/err.2024.4.

Miner, A. S., Shah, N., Bullock, K. D., Arnow, B. A., Bailenson, J., & Hancock, J. (2019). Key Considerations for Incorporating Conversational AI in Psychotherapy. *Frontiers in psychiatry*, *10*, 746. https://doi.org/10.3389/fpsyt.2019.00746

Mokeke, I., Kim, S., & Choi, Y.-S. (2024). Adhering to patient safety standards: A healthcare imperative for the 21st century. Medicine and Clinical Science, 6, Article 1127. https://doi.org/10.33425/2690-5191.1127

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025, June). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 599–627). https://doi.org/10.48550/arXiv.2504.18412

Ni, Y., & Jia, F. (2025). A scoping review of AI-driven digital interventions in mental health care: Mapping applications across screening, support, monitoring, prevention, and clinical education. Healthcare,13(10), 1205. https://doi.org/10.3390/healthcare13101205

OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD Publishing. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Omar, M., Sorin, V., Agbareia, R., Apakama, D., Soroush, A., Sakhuja, A., Freeman, R., Horowitz, C., Richardson, L., Nadkarni, G., & Klang, E. (2024). *Evaluating and addressing demographic disparities in medical large language models: A systematic review*. International Journal for Equity in Health, 24, 419. https://doi.org/10.1186/s12939-025-02419-0

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with artificial intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, Article 100099. https://doi.org/10.1016/j.glmedi.2024.100099

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

Perneger, T. V. (2005). The Swiss cheese model of safety incidents: Are there holes in the metaphor? BMC Health Services Research, 5, 71. https://doi.org/10.1186/1472-6963-5-71

Pinsent Masons. (2024, February 13). A guide to high-risk AI systems under the EU AI Act [White paper]. https://www.pinsentmasons.com/out-law/guides/guide-to-high-risk-ai-systems-under-the-eu-ai-act

Reason, J. (2000). Human error: Models and management. BMJ, 320(7237), 768–770. https://doi.org/10.1136/bmj.320.7237.768

Rollwage, M., Habicht, J., Juechems, K., Carrington, B., Viswanathan, S., Stylianou, M., Hauser, T. U., & Harper, R. (2023). Using conversational AI to facilitate mental health assessments and improve clinical efficiency within psychotherapy services: Real-world observational study. JMIR AI, 2, e44358. https://doi.org/10.2196/44358

Sarkar, S., et al. (2023). A review of the explainability and safety of conversational agents for mental health. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2023.1229805

Sedlakova, J., & Trachsel, M. (2023). Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?. *The American journal of bioethics : AJOB*, *23*(5), 4–13. https://doi.org/10.1080/15265161.2022.2048739

Seshia, S., Young, G., Makhinson, M., Makhinson, M., Smith, P., Stobart, K., & Croskerry, P. (2017). Gating the holes in the Swiss cheese (part I): Expanding professor Reason's model for patient safety. Journal of Evaluation in Clinical Practice, 24, 187 - 197. https://doi.org/10.1111/jep.12847

Shabani, T., Jerie, S., & Shabani, T. (2023). A comprehensive review of the Swiss cheese model in risk management. Safety in Extreme Environments, 6, 43-57. https://doi.org/10.1007/s42797-023-00091-7.

Sharif, L., Almabadi, R., Alahmari, A., Alqurashi, F., Alsahafi, F., Qusti, S., Akash, W., Mahsoon, A., Poudel, D. B., Sharif, K., & Wright, R. (2025). Perceptions of mental health professionals towards artificial intelligence in mental healthcare: A cross-sectional study. Frontiers in Psychiatry, 16, Article 1601456. https://doi.org/10.3389/fpsyt.2025.1601456

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, *5*(1), 46-57. https://doi.org/10.48550/arXiv.2203.15144

Siddals, S., Torous, J., & Coxon, A. (2024). “It happened to be the perfect thing”: Experiences of generative AI chatbots for mental health. npj Mental Health Research, 3, 48. https://doi.org/10.1038/s44184-024-00097-4

Tavory, T. (2024). *Regulating AI in mental health: Ethics of care perspective*. JMIR Mental Health, 11, e58493. https://doi.org/10.2196/58493

TIME. (2025, March 18). Parents allege ChatGPT responsible for son’s death by suicide. TIME Magazine. https://time.com/7312484/chatgpt-openai-suicide-lawsuit

TÜV AI Lab. (2023). AI, MDR & AI Act – White paper. TÜV AI Lab. https://www.tuev-lab.ai/fileadmin/user_upload/TUEV_AI_Lab_Whitepaper_MDR_AIAct_EN.pdf

TÜV SÜD. (n.d.). AI regulations in healthcare: What EU AI Act 2024/1689 means for medical device manufacturers [White paper]. https://www.tuvsud.com/en/knowledge-hub/white-papers/ai-regulations-in-healthcare-what-the-eu-ai-act-2024-1689-means-for-medical-device-manufacturers

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. Canadian Journal of Psychiatry, 64(7), 456–464. https://doi.org/10.1177/0706743719828977

van der Schyff, E. L., Ridout, B., Amon, K. L., Forsyth, R., & Campbell, A. J. (2023). Providing Self-Led Mental Health Support Through an Artificial Intelligence-Powered Chat Bot (Leora) to Meet the Demand of Mental Health Care. *Journal of medical Internet research*, *25*, e46448. https://doi.org/10.2196/46448

Van Gerven, E., Bruyneel, L., Panella, M., Euwema, M., Sermeus, W., & Vanhaecht, K. (2016). Psychological impact and recovery after involvement in a patient safety incident: a repeated measures analysis. *BMJ Open*, 6. https://doi.org/10.1136/bmjopen-2016-011403.

Varastehnezhad, A., Tavasoli, R., Elyasi, S., LotfiNia, M., & Farbeh, H. (2025). AI in Mental Health: Emotional and Sentiment Analysis of Large Language Models' Responses to Depression, Anxiety, and Stress Queries. ArXiv, abs/2508.11285. 10.48550/arXiv.2508.11285

Wang, X., Zhou, Y., & Zhou, G. (2025). The Application and Ethical Implication of Generative AI in Mental Health: Systematic Review. *JMIR mental health*, *12*, e70610. https://doi.org/10.2196/70610

Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2025). Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. PLOS Digital Health, 4(4), e0000810. https://doi.org/10.1371/journal.pdig.0000810

Widder, D., & Nafus, D. (2022). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10. https://doi.org/10.1177/20539517231177620.

World Health Organization Regional Office for Europe. (2025, November 19). *Artificial intelligence is reshaping health systems: State of readiness across the WHO European Region* (WHO/EURO:2025-12707-52481-81028). World Health Organization. https://www.who.int/europe/publications/i/item/WHO-EURO-2025-12707-52481-81028

Wiegmann, D. A., Wood, L. J., Cohen, T. N., & Shappell, S. A. (2022). Understanding the "Swiss Cheese Model" and its application to patient safety. Journal of Patient Safety, 18(2), 119–123. https://doi.org/10.1097/PTS.0000000000000810

Appendix

Appendix A

Overview Interviewees

ID	Stakeholder Group	Professional Role	Professional Experience	Country	Age Group	Gender
C01	Clinicians (C)	Psychologist & Psychological Psychotherapist in Training (acute psychiatry)	Clinical experience in acute psychiatry and crisis ward; work with suicidality, psychosis, schizoaffective disorders	Germany	30–35	female
C02	Clinicians (C)	Coach (Systemic Consultation; non-clinical)	Professional experience in systemic coaching with vulnerable groups; online coaching formats; work with young adults in stressful or transitional situations	Germany	25–30	female
C03	Clinicians (C)	Physician & Medical AI Researcher	Clinical care experience combined with interdisciplinary research on AI systems in healthcare	Germany	30-35	female
C04	Clinicians (C)	Licensed Psychological Psychotherapist	~20+ years of clinical experience in outpatient psychotherapy	Germany	Not disclosed	female
IE01	Industry Experts (IE)	Chief Commercial Officer, Digital Therapeutics	Senior industry executive; 20+ years in market access, commercialisation, and development of regulated digital therapeutics	United States	55–60	male
IE02	Industry Experts (IE)	Developer / Technical Expert (AI/ML)	Early-career technical expert; several years of academic experience in AI/ML, NLP, bias & fairness research	United Kingdom	25-30	female
IE03	Industry Experts (IE)	Researcher, LLM Behaviour & Safety	3+ years in LLM alignment, behaviour shaping, and safety research at a frontier model lab	United Kingdom	25-30	male
IE04	Industry Experts (IE)	Senior Strategist, Digital Therapeutics	7+ years in mental-health digital health (UX, clinical product development, regulatory coordination)	Germany	35-40	female
P01	Patients (P)	Student / Young Adult (mental-health service user)	Lived experience with depression, mood instability, and ongoing psychotherapy; repeated use of AI chatbots during vulnerable phases	Germany	20-25	female
P02	Patients (P)	Young Adult (mental-health service user)	Lived experience with hypochondria, panic attacks, and health anxiety; frequent use of AI chatbots for reassurance and panic regulation	Germany	25-30	female
P03	Patients (P)	Young Adult (mental-health service user)	Lived experience with low mood, rumination, loneliness; use of AI chatbots for emotional regulation and cognitive structuring	Portugal	25-30	male
R01	Regulators (R)	Legal expert in EU AI governance & digital regulation	Lawyer specialising in EU digital regulation & AI governance	Germany	45-50	female
R02	Regulators (R)	Legal expert in EU AI governance & digital regulation	Early-career legal expert; completed legal studies and LL.M.; academic and regulatory-related experience in AI governance, GDPR, and high-risk technology oversight	United Kindom	25-30	female
R03	Regulators (R)	Legal scholar specialising in EU law, patients' rights & AI regulation	Early- to mid-career academic; several years of research experience in EU regulation, mental-health law, and technology governance	Italy	25-30	male

Appendix B

Results - Structured by Topics of IQ Questions - Clinicians

		Clinician			
Sub-Question | Focus	**Main-Topic**	**C01**	**C02**	**C03**	**C04**
SQ1 | Responsibilities & Controls	**Distributed Responsibility Across Actors** *"Patients should not use AI for diagnoses or as a substitute for therapy." (C01)*	shared roles; patient limits	roles defined; user limits	role distribution	shared responsibility
	Need for Clear Boundaries and Transparency *"Platforms must be transparent, provide clear warnings, and communicate where their limits are." (C02)*	reliability limits	warnings; boundaries	clarity; explanations	safety rails
	Misaligned Expectations Around AI Use *"We can't expect them to carry the main safety burden." (C04)*	self-diagnosis; expectations	boundary confusion	user knowledge limits	vulnerable users
SQ2 | Vulnerabilities	**High-Risk Situations** *"Acute suicidality." (C01)*	acute suicidality; delusional reinforcement	severe distress episodes	critical symptoms overlooked	misreading stress vs danger
	Overreliance & Misreading *"Formulates its recommendations as if it were completely certain." (C03)*	unquestioned trust; overreliance	pseudo-therapy risk	authoritative tone; overconfidence	misclassification of signals
	Pseudo-Empathy & Emotional Risk *"Pseudo-empathy." (C04)*	false reassurance; no nuance	emotional misattunement	misplaced comfort; illusion of care	comforting tone without grounding
	Lack of Evaluation Capacity *"My current role is none." (C01)*	no evaluation ability	unclear capabilities; uncertainty	limited insight into systems	information gap; fast market
SQ3 | Cross-Layer Coordination (Interdependencies)	**Limited Collaboration** *Built with "very little clinical input." (C04)*	limited co-design	little therapist input	rare structured exchange	minimal clinical involvement
	Unclear Responsibilities *"Who is responsible? The app provider, the model provider, the app store, or the therapist?" (C04)*	unclear escalation roles	responsibility confusion	uncertain accountability	unclear incident roles
	Need for Structured Clinical Involvement *Qualified professionals should "review assessments and intervene." (C03)*	mandatory clinical role	importance of boundaries	review and intervention	clinical expertise needed
SQ4 | Safeguards & Governance	**Supplement Not Replacement** *"AI should only be used as a supplement, not as a replacement for professional treatment." (C01)*	supplement only; psychoeducation; journaling; skills	not a therapeutic service; label clearly	cannot replace empathy; nonverbal loss	bridge on waiting list; skills between sessions; human-led for severe cases
	Clear Limits and Crisis Protocols *"Clear criteria for crisis detection and appropriate escalation pathways." (C01)*	crisis sensitivity; escalation criteria	crisis warnings; critical keyword detection; clear recommendations	harder thresholds in mental health; individual patterns	robust crisis protocols; red lines (no self-harm promotion, no erotic/romantic chat); monitoring overuse/isolation
	Professional Oversight *"Qualified professionals should be involved, and system accuracy should be reviewed regularly." (C03)*	certification; clinical studies; continuous monitoring; data protection	design principles under clear governance; transparent data use	regular accuracy checks; professional involvement	staged access via therapist or health system; defined handover point; supervised use

Results - Structured by Topics of IQ Questions - Industry Experts

		Industry Experts			
Sub-Question | Focus	**Main-Topic**	**IE01**	**IE02**	**IE03**	**IE04**
SQ1 | Responsibilities & Controls	**Defined Roles Across Actors** *"Each group has a clear lane." (IE01)*	**regulator roles; developer duties**	**testing responsibility**	**shared boundaries**	**scope definition**
	Boundary-Setting and Transparency *"Tone is sometimes the biggest culprit." (IE03)*	**expectation management**	**data rules**	**tone limits; guardrails**	**avoid therapy tone**
	Responsibility Limits for Developers *"Once a model leaves the lab… risk shifts." (IE03)*	**cannot solve everything**	**supervised compliance**	**downstream limits**	**regulatory gaps not covered**
SQ2 | Vulnerabilities	**Technical Weak Points** *Underweighted edge cases "produce unreliable outputs." (IE02)*	**missed bugs; small team constraints**	**dataset bias; incomplete coverage**	**classifier misses short signals**	**limits of system reliability**
	Distress-Driven User Vulnerability *"Distress changes everything." (IE03)*	**vulnerability under overload**	**fragile states increase risk**	**heightened suggestibility**	**emotional overwhelm patterns**
	Overconfidence & False Empathy *"False empathy." (IE03)*	**tone conveys authority**	**unclear emotional cues**	**warm tone misleads**	**trust projected onto system**
	Safeguard Limitations *Users may "project more competence" onto systems. (IE04)*	**post-market gaps; missed risks**	**incomplete testing cycles**	**boundary blurring in use**	**expectation mismatches**
SQ3 | Cross-Layer Coordination (Interdependencies)	**Early Alignment** *Built trust through "scientific rigor." (IE01)*	**advisory boards; evidence**	**early requirement sync**	**alignment on safety goals**	**stakeholder coordination**
	Communication Gaps *Developers "left to follow the project specifications on their own." (IE02)*	**unclear expectations**	**unclear data rules**	**missing shared language**	**coordination gaps**
	Responsibility Diffusion *Misaligned prompts can make users "fall into the gap." (IE03)*	**unclear handovers**	**mixed accountability**	**model vs app confusion**	**split responsibilities**
	Differing Incentives *Teams must "hold the centre." (IE04)*	**clinical vs tech priorities**	**regulatory vs delivery speed**	**emotional risk not aligned**	**diverging team goals**
SQ4 | Safeguards & Governance	**Clinical-Grade Constraints** *"Conversational AI must be required to have similar clinical evidence as expert-system-based DTx of the first generation." (IE01)*	**guideline-constrained models; prevent hallucinations**	**sensitive-data awareness; avoid unsafe uses**	**prevent improvisational empathy; narrow crisis role**	**strict behavioural constraints; no improvisation in dialogue**
	Evidence and Oversight *"Nothing is released without evidence, clinical sign-off, and regulatory approval." (IE04)*	**ethical reviews; RCTs; evidence-driven decisions; continuous-evidence frameworks**	**employee training; security certifications; user briefings on data use**	**audit trails; risk-tiered oversight**	**responsibility over speed; internal iteration before release**
	Risk-Sensitive Technical Controls *"The safest answer is narrow: detect, de-escalate, hand off." (IE03)*	**guardrails around patient safety**	**technical controls supported by frameworks**	**crisis indicator detection; risk-sensitive decoding; external governors**	**concern about agents and companions; capability–regulation gap**

Results - Structured by Topics of IQ Questions - Patients

		Patients		
Sub-Question | Focus	**Main-Topic**	**P01**	**P02**	**P03**
SQ1 | Responsibilities & Controls	**Limited Responsibility for Users** *"You can't always judge what's safe or healthy." (P01)*	**limited user role**	**panic; unclear judgment**	**distress limits**
	Primary Responsibility for Developers and Regulators *"It's unrealistic to expect people to self-regulate perfectly when they're in distress." (P03)*	**developer/regulator duty**	**experts should decide**	**define limits**
	Need for Clear Guidance in Crisis Moments *"This is something you shouldn't handle alone, please reach out to someone." (P03)*	**crisis prompts**	**direct guidance**	**clearer signals**
SQ2 | Vulnerabilities	**Generic or Misattuned Responses** *Advice felt like a "checklist" when "very desperate." (P01)*	**generic advice; mismatch**	**impersonal suggestions**	**emotional tone mismatch**
	Misreading Emotional Intensity *Could not detect being "on the verge of doing something impulsive." (P01)*	**severity misread; urgency missed**	**panic cues overlooked**	**nuance not detected**
	Overanalysis & Diagnostic Drift *Received explanations of disorders after saying "I feel empty today." (P03)*	**unwanted analysis**	**too much depth**	**diagnostic overreach**
	Reinforcing Rumination *"Doesn't always recognize when I'm stuck in rumination." (P03)*	**circular reassurance**	**worry loop reinforced**	**rumination not interrupted**
SQ3 | Cross-Layer Coordination (Interdependencies)	**Partial Trust** *Real relationships offer something "the AI lacks." (P01) Quote*	**limited trust**	**info but not guidance**	**supportive but limited**
	Using AI Before Humans *Use it "for a first impression." (P02)*	**AI before contacting people**	**first step use**	**initial support**
	Delayed Help Seeking *A "safe middle ground" delaying contact. (P03)*	**self-isolation episodes**	**repeated reassurance**	**delayed escalation**
SQ4 | Safeguards & Governance	**Crisis Differentiation** *"If language suggesting deep despair, hopelessness or self-harm appears, the AI should clearly state that it cannot replace emergency help." (P01)*	**distinguish info vs crisis; auto-referral to hotlines**	**threshold-based referral to real support**	**detect spiralling; switch to grounding responses**
	Clear Boundaries *"I'm not a therapist, but here are some basic strategies." (P03,*	**repeated reminders not a professional; cannot replace therapy**	**avoid heavy, dramatic phrasing; avoid worst-case focus**	**explicit non-therapist framing; transparency about data use**
	Harm-Avoidance Rules *"It should never give instructions for self-harm, minimise or compare feelings, or shift all responsibility back onto the user." (P01)*	**no self-harm instructions; no minimising or romanticising suffering; encourage involving real people**	**avoid dramatizing answers; protect against dangerous topics**	**no downplaying risk; no harmful coping suggestions; no guessing or labelling diagnoses**

Results - Structured by Topics of IQ Questions - Regulators

		Regulators		
Sub-Question | Focus	**Main-Topic**	**R01**	**R02**	**R03**
SQ1 | Responsibilities & Controls	**Setting Standards and Normative Requirements** *Rights like privacy must be "operationalised." (R03)*	**rules; standards**	**guidance; oversight**	**high-risk criteria**
	Developers Implement Legal and Technical Controls *Developers should not hide behind the "black box" argument. (R03)*	**risk mgmt; governance**	**testing; monitoring**	**bias tests; limits**
	Clinicians Maintain Clinical Judgment *"Should not outsource all decision-making to the system."*	**clinical autonomy**	**clinical autonomy**	**hybrid use**
	Limited Responsibilities for End Users *"Too much responsibility cannot be placed on them." (R02)*	user autonomy	user limits	user rights
SQ2 | Vulnerabilities	**Psychological Harm Limits** *Avoid "aggravat[ing] existing power imbalances." (R01)*	**emotional/cognitive harm**	**dependency and trust risks**	**subtle relational effects**
	Medical–Non-Medical Grey Zone *Boundaries "complicate" decisions. (R01)*	**unclear role boundaries**	**ambiguity in advice**	**shifting system function**
	Legal Abstraction & Pace *Requirements are "quite abstract." (R03)*	**broad legal framing**	**pacing problem; slow clarity**	**abstract obligations**
	Missing Interdisciplinary Input *Input "not always integrated." (R01)*	**missing psychology expertise**	**limited clinical insight**	**siloed areas of knowledge**
SQ3 | Cross-Layer Coordination (Interdependencies)	**Fragmented Collaboration** *"Legal expectations, technical implementation, and clinical realities do not always meet in a coherent way." (R01)*	**inconsistent collaboration**	**irregular stakeholder input**	**siloed processes**
	Clinical–Technical Disconnect *"No formal mechanism guaranteeing clinical involvement." (R02)*	**limited psychological input**	**absent clinical channel**	**inadequate expertise flow**
	Accountability Gaps *Accountability becomes "unclear" when providers differ. (R02)*	**unclear responsibility**	**multi-actor ambiguity**	**overlapping obligations**
	Transparency and Resource Limits *"Structural knowledge gap." (R02)*	**limited data access**	**scarce regulatory resources**	**unclear risk interpretations**
	Jurisdiction Differences *Differing national and international requirements. (R01, R03)*	**cross-country variation**	**inconsistent safeguards**	**divergent regulations**
SQ4 | Safeguards & Governance	**Liability and Transparency** *"Liability regimes should reflect that psychological and emotional harms can be just as serious as physical harms." (R01)*	**clear liability rules; documentation duties; fundamental-rights impact assessments**	**safety baseline; classification; traceability**	**clarify responsibility in clinical use; liability for scaled deployment**
	Independent Oversight *"Independent certification and auditing, not only for technical robustness but also for psychological safety." (R01)*	**independent certification; audits of psychological safety**	**continuous monitoring; incident reporting; threshold interventions**	**interdisciplinary audits; stronger enforcement structures; specialised units**
	Psychological Safety Standards *"I would define psychological safety explicitly and embed it directly into conformity assessments." (R02)*	**certificates including emotional risk; vulnerability and fundamental rights**	**explicit psychological safety; high-risk mental-health tools**	**mental-health-specific concerns; naming psychotherapy; stronger rules on explainability, data and oversight; high-risk use cases (profiling, monitoring, triage)**

Appendix C

Discussion - Interpretation

Sub-Question | Focus	Main-Topic	Similar Patterns	Contradictions/ Tensions	Swiss Cheese Model		Literature Reference
				Layers	Holes	
SQ1 | Responsibilities & Controls	**Distributed Responsibility Across Actors**	**Strong cross-stakeholder consensus that patients should not carry primary safety responsibility.**	**Regulators frame responsibility as rights-based and abstract, while developers treat it as design- and lifecycle-based, and clinicians as situational and judgment-based.**	**Regulatory layer (baseline guardrails)**	**Responsibility treated as transferable instead of relational**	**Swiss Cheese Model (Reason)**
	"Patients should not use AI for diagnoses or as a substitute for therapy." (C01)	*Shared recognition that safety is distributed across layers, not owned by a single actor.*	*Tension between standardisation (rules, disclaimers) and clinical discretion.*	*Developer layer (operational safety-by-design)*	*Boundary-setting misaligned with clinical reality*	*Distributed responsibility in socio-technical systems*
	Need for Clear Boundaries and Transparency	**Agreement on the need for clear boundaries to prevent role confusion (AI ≠ therapist).**	**Implicit assumption by each layer that another layer will intercept failures.**	**Clinical layer (contextual judgment & escalation)**	**Lack of end-to-end accountability across layers**	**Patient vulnerability & impaired judgment under distress**
	"Platforms must be transparent, provide clear warnings, and communicate where their limits are." (C02)			Patient layer (protected, not responsible)		
	Misaligned Expectations Around AI Use	**Convergence on transparency and guidance as protective mechanisms.**				
	"We can't expect them to carry the main safety burden." (C04)					
SQ2 | Vulnerabilities	**High-Risk Situations**	**Agreement that risk is systemic, not caused by single failures.**	**Developers focus on technical limits, clinicians on therapeutic miscalibration.**	**Technical layer (risk detection, tone, escalation logic)**	**Failure to differentiate low-risk vs high-risk states**	**Pseudo-empathy & emotional overvalidation**
	"Acute suicidality." (C01)	*Shared concern around high-risk situations (e.g., suicidality).*	*Regulators under-specify psychological escalation as a safety category.*	*Clinical layer (absence of evaluation frameworks)*	*Empathic tone without clinical accountability*	*Overreliance on digital mental health tools*
	Overreliance & Misreading	**Recognition of pseudo-empathy as a recurrent failure mode.**	**Patients experience harm experientially, while upstream layers frame it structurally.**	**Regulatory layer (oversight blind spots)**	**Continuous availability without enforced stopping rules**	**Vulnerability amplification in socio-technical systems**
	"Formulates its recommendations as if it were completely certain." (C03)	*Consensus that users least able to compensate are most exposed.*		*Patient layer (overreliance under distress)*	*Oversight gaps across deployment contexts*	
	Pseudo-Empathy & Emotional Risk					
	"Pseudo-empathy." (C04)					
	Lack of Evaluation Capacity					
	"My current role is none." (C01)					
SQ3 | Cross-Layer Coordination (Interdependencies)	**Limited Collaboration**	**Broad agreement that interfaces, not components, determine safety.**	**Clinicians see misalignment as loss of therapeutic calibration.**	**Developer–clinical interface**	**Unclear handovers between layers**	**Interface failures in safety-critical systems**
	Built with "very little clinical input." (C04)	*Shared recognition that late clinical involvement increases risk.*	*Developers frame it as innovation-cycle constraints.*	*Regulatory–developer interface*	*Semantic misalignment ("support" vs "therapy")*	*Responsibility diffusion in modular architectures*
	Unclear Responsibilities	**Consensus that modular AI ecosystems diffuse responsibility.**	**Regulators interpret it as governance gaps, not design failures.**	**AI–patient interface**	**Escalation based on assumption rather than enforcement**	**Handover failures & coordination risk**
	"Who is responsible? The app provider, the model provider, the app store, or the therapist?" (C04)	*Agreement that escalation pathways are critical but weakly enforced.*	*Normative agreement on escalation vs. operational diffusion of responsibility.*			
	Need for Structured Clinical Involvement					
	Qualified professionals should "review assessments and intervene." (C03)					
SQ4 | Safeguards & Governance	**Supplement Not Replacement**	**Agreement that AI must supplement, not replace human care.**	**Industry/regulators fear over-constraint harming innovation.**	**Design layer (constraint-based safeguards)**	**Implicit safeguards relying on user discretion**	**Safety-by-design**
	"AI should only be used as a supplement, not as a replacement for professional treatment." (C01)	*Shared emphasis on clear limits, escalation, and stopping rules.*	*Clinicians/patients prioritise protection in high-risk scenarios.*	*Governance layer (liability, monitoring)*	*Fragmented governance across abstraction levels*	*Risk-tiered governance*
	Clear Limits and Crisis Protocols	**Convergence on need for visible, explicit safeguards.**	**Divergence on where safeguards should primarily reside (design vs oversight).**	**Clinical layer (professional oversight)**	**Static rules applied to dynamic systems**	**Proportionality in digital health regulation**
	"Clear criteria for crisis detection and appropriate escalation pathways." (C01)	*Recognition that safeguards must operate across layers.*		*Patient layer (experiential safety cues)*		
	Professional Oversight					
	"Qualified professionals should be involved, and system accuracy should be reviewed regularly." (C03)					

Appendix D

Interview Transcripts (Availability Notice)

Full verbatim interview transcripts were prepared as part of this dissertation's research process. However, due to the limited page allowance, they are not included in the appendix of this submitted document. The transcripts are stored separately and can be made available upon request at any time.