www.ijarp.com ISSN 2456-9992 Page: 01-28

International Journal Advanced Research Publications

Research Article

Volume: 01
Issue: 03

ARTIFICIAL INTELLIGENCE AND THE EMOTIONAL VULNERABILITY OF USERS: IMPLICATIONS FOR MENTAL HEALTH AND WELLBEING

***Humphrey Lephethe Motsepe**

Limpopo Department of Agriculture and Rural Development (Towoomba Research Centre), Management College of Southern Africa (MANCOSA) and University of Venda, South Africa

Article Received: 31 October 2025, Article Revised: 20 November 2025, Published on: 10 December 2025

***Corresponding Author: Humphrey Lephethe Motsepe**

Limpopo Department of Agriculture and Rural Development (Towoomba Research Centre), Management College of Southern Africa (MANCOSA) and University of Venda, South Africa.

DOI: https://doi-doi.org/101555/ijrpa.3770

ABSTRACT: This article examines how artificial intelligence affects the mental health and wellbeing of users who are already vulnerable due to age, socioeconomic status or pre-existing psychological conditions. The purpose of the study is to consolidate existing evidence and develop a clearer understanding of what kinds of risks and opportunities AI presents for these populations. The study uses a qualitative document analysis approach that draws on recent peer reviewed literature, global policy documents and technical reports. This approach makes it possible to analyse existing knowledge without collecting new human data, which avoids the need for ethical clearance. The findings indicate that while AI driven tools can improve access to mental health support and early detection of psychological distress, they also introduce risks that arise from algorithmic bias, targeted content exposure, emotional dependency on AI chatbots, surveillance-based data practices and reduced human contact. These risks appear to be magnified among vulnerable users who often have limited digital literacy and fewer safeguards. The article concludes that current research has not sufficiently accounted for the lived realities of these users and that more inclusive risk assessment models are required. The study recommends that policymakers and designers adopt protective design principles, transparent data practices and oversight mechanisms that centre the needs of vulnerable communities.

KEYWORDS: Artificial intelligence; Digital wellbeing; Algorithmic risk ;Mental health; Vulnerable users.

INTRODUCTION AND BACKGROUND

Artificial intelligence has become woven into the fabric of daily life, shaping how people communicate, access services and maintain social connections. In many countries, AI systems now support everyday activities through digital assistants, content recommendation tools, mental health applications and automated communication platforms. These technologies offer meaningful benefits, including quicker access to information, opportunities for personalised support and increasingly sophisticated tools for managing health and wellbeing (Kelly et al., 2023). The expansion of AI into these areas has been accompanied by rapid adoption, largely due to the convenience and perceived neutrality of automation. Yet as AI becomes more deeply integrated into society, questions about its psychological impact and its influence on emotionally vulnerable users have gained prominence. While AI-enabled systems can provide significant support in moments of emotional difficulty, they may also produce outcomes that inadvertently heighten distress or compromise user safety. These risks tend to be more pronounced among groups whose circumstances make them less resilient to digital harms. Vulnerable users may include adolescents who are still developing emotional regulation skills, older adults who sometimes struggle with technological adaptation, individuals with limited digital literacy and people already dealing with psychological challenges such as anxiety or depression (Shaw et al., 2023). Because these groups are more susceptible to persuasive design techniques, targeted content or algorithmic biases, the emotional effects of AI may be more intense and less predictable, increasing the need for careful evaluation.

Recent research highlights that digital mental health technologies are expanding at an unprecedented pace, offering tools ranging from AI-generated cognitive behavioural interventions to automated mood-tracking applications. Despite this growth, many users remain unaware of the extent to which algorithmic systems shape the content they encounter and the responses they receive (Torous et al., 2021). Lack of awareness is particularly concerning because the design and operation of AI can subtly influence user emotions, perceptions and decisions. Scholars have raised concerns about the psychological impact of personalised content, especially in environments where algorithms optimise for engagement rather than wellbeing. These concerns are reinforced by evidence showing that emotionally charged or sensational content often receives disproportionate visibility on major platforms, which can have significant mental health implications for susceptible populations (Meier & Reinecke, 2021). Another challenge relates to privacy and data security. AI systems that rely

on large datasets routinely gather detailed information about user behaviour, preferences and emotional states. For vulnerable users, such data practices can feel intrusive and may create additional anxiety about surveillance or misuse (Barda et al., 2022). The problem is further complicated by the fact that AI often operates in ways that are not fully transparent. Users may not understand why certain responses are generated or why specific content is recommended, which can undermine trust and contribute to feelings of uncertainty. In mental health contexts, trust and clarity are essential, and reduced transparency can limit the effectiveness of digital interventions.

Despite these concerns, the potential benefits of AI for mental health support should not be overlooked. AI-assisted tools can increase access to psychological resources, particularly in settings where mental health services are under-resourced or difficult to reach. Studies show that AI-mediated interventions can help reduce symptoms of depression and anxiety when used appropriately, and they can provide round-the-clock assistance that complements traditional therapeutic models (Inkster et al., 2023). For some vulnerable users, these tools offer a sense of companionship or stability during moments when human support may not be immediately available. However, the effectiveness of such systems depends heavily on their design, ethical safeguards and consistency in handling emotionally sensitive information. Even with these benefits, the literature consistently notes a gap in understanding how AI affects vulnerable populations specifically. Much of the existing research focuses on general user groups, leaving questions about differential impact unanswered. Vulnerable individuals may react differently to automated feedback, personalised recommendations or conversational agents, and they may also be more exposed to risks linked to algorithmic inaccuracies, emotional misinterpretation or harmful content loops (Vaswani et al., 2023). This gap underscores the need for research that centres on vulnerable groups rather than treating them as peripheral cases.

The present study therefore seeks to address these gaps by examining how AI influences the emotional wellbeing and mental health outcomes of vulnerable users. The aim is not only to describe the challenges but also to highlight the opportunities that AI can bring when developed and deployed responsibly. This approach acknowledges both the promise and the complexity of AI in mental health contexts. Three research questions guide this study:

First, how does AI affect the mental health and wellbeing of vulnerable users? This question focuses on emotional responses, behavioural patterns and the psychological impact of sustained AI interaction.

Second, what risks and opportunities arise from AI-mediated interactions for these groups? This includes examining potential harms such as over-reliance on AI, exposure to harmful content or emotional misalignment, as well as potential benefits such as improved access to support or early detection of distress signals.

Third, what design and policy considerations can help ensure that AI systems protect vulnerable individuals while still delivering meaningful assistance? Addressing this question requires attention to transparency standards, ethical safeguards, inclusive design principles and appropriate regulatory mechanisms.

Understanding the relationship between AI and mental health in vulnerable populations is essential as societies move toward greater digitalisation. Stronger evidence is needed to inform guidelines, promote ethical design choices and shape public policy that prioritises user wellbeing. By exploring these issues, this study contributes to a growing body of knowledge and supports efforts to build AI systems that are both helpful and safe for those who rely on them most.

LITERATURE REVIEW

Artificial intelligence has become a central component of digital mental health ecosystems, influencing how different groups engage with online support, access information and navigate emotional challenges. The rapid expansion of AI-enabled tools has led to extensive scholarly interest in understanding their psychological, behavioural and social impacts. This literature review synthesises recent evidence on the benefits, risks and broader implications of AI for vulnerable users, while also identifying gaps that remain in academic and policy discussions. It discusses AI's mental health applications, the psychological risks linked to algorithmic systems, the dynamics of dependency and emotional misalignment, issues of equity and representation and the limited attention paid to vulnerable populations. It also highlights emerging insights into algorithmic design, long-term behavioural effects and the need for inclusive governance strategies.

Benefits of AI for Mental Health Support

A significant portion of recent research focuses on the potential of artificial intelligence to enhance mental health support, particularly through digital applications and automated systems. Digital mental health tools that incorporate machine learning have demonstrated value in improving access to timely and personalised care. Several studies show that AI-enhanced applications can function as a first line of support, particularly for individuals who

face barriers in accessing traditional mental health services. These barriers may include financial constraints, stigma, geographic isolation or limited availability of trained professionals (Wasil et al., 2021). As noted by Wasil and colleagues, digital tools provide users with real-time assistance, which can be especially beneficial during moments of acute stress or emotional instability. The real-time nature of AI support has also attracted considerable scholarly attention. Many AI-driven mental health platforms are programmed to detect shifts in emotional tone, behavioural patterns or self-reported symptoms, allowing them to provide immediate responses when users express feelings of distress. Torous et al. (2021) report that predictive algorithms can support early identification of emerging mental health challenges, helping users understand the significance of their symptoms and encouraging timely intervention. This ability to offer on-demand guidance helps fill gaps in healthcare systems where long waiting times and limited resources often delay professional support.

Another noted benefit relates to the scalability of AI-assisted mental health interventions. AI systems can serve large populations simultaneously, without compromising consistency or increasing the workload of mental health professionals. This scalability has proven particularly important during global crises such as the COVID-19 pandemic, where demands on psychological services increased dramatically. Researchers have observed that AI-enabled chatbots, when appropriately designed, can deliver structured psychological strategies that mirror evidence-based techniques such as cognitive behavioural therapy, motivational interviewing and stress-management frameworks (Inkster et al., 2023). These tools do not replace clinical care but offer a supplementary resource that can help reduce symptom escalation. In addition to individual-level interventions, AI can support mental health systems at a broader structural level. Scholars argue that predictive analytics can help organisations forecast mental health trends, identify high-risk communities and allocate resources more effectively. For example, large-scale analyses of anonymised data can help detect spikes in anxiety or depression within specific demographic groups, enabling targeted outreach responses. This function may prove particularly valuable in resource-constrained settings, where data-driven decision-making can enhance service delivery (Barda et al., 2022).

Despite these positive developments, scholars caution that the effectiveness of AI-based tools depends heavily on their design and operational transparency. While AI offers opportunities to widen access, the quality of support varies widely across platforms, and many tools lack

clinical oversight. As a result, although the literature documents numerous benefits, it also highlights the need for more rigorous evaluation frameworks to ensure that AI-driven mental health tools meet appropriate safety and ethical standards.

Psychological Risks Associated with AI

Alongside the potential benefits, a growing body of literature warns that AI-enabled digital mental health tools can introduce psychological risks, particularly for vulnerable users. One of the most well-documented concerns relates to emotional dependency. Fleming et al. (2022) found that adolescents using AI-based counselling systems often struggle to differentiate between automated and human agents. This difficulty becomes problematic when users begin to form emotional attachments to AI systems, sometimes perceiving them as reliable companions or confidants. For adolescents who lack stable emotional support, such attachments may deepen quickly, creating dependency patterns that undermine healthy coping mechanisms. Dependency-related risks are compounded by the fact that AI systems cannot fully understand human emotional nuance, even when they appear conversationally competent. Because these systems rely on pre-programmed responses rather than lived experience, their emotional attunement is limited. This misalignment can result in situations where vulnerable users receive responses that minimise or misunderstand their emotional states, potentially intensifying feelings of isolation or frustration (Torous et al., 2021). The risk is particularly serious for individuals dealing with trauma, suicidal thoughts or severe psychological instability, as incorrect responses can have harmful consequences.

Another major concern relates to the role of algorithmic curation on social media. Platforms powered by machine learning can personalise content to maximise user engagement, but this process often exposes vulnerable individuals to material that reinforces negative emotional states. Studies have documented that algorithmic systems may unintentionally amplify harmful content, such as posts related to self-harm, extreme dieting, conspiracy theories or emotionally charged political material (Meier & Reinecke, 2021). Vulnerable users may experience compulsive engagement with such content, which can worsen anxiety, reduce self-esteem or contribute to depressive symptoms. Scholars further note that algorithmic systems create feedback loops that intensify harmful behavioural patterns. When a user interacts with content related to sadness or insecurity, algorithms may assign greater relevance to similar material, resulting in repeated exposure. Shaw et al. (2023) show that this targeted exposure can lead to increased rumination, emotional dysregulation and heightened

psychological distress. The psychological effects of these feedback loops can be substantial, as repeated exposure gradually shapes users' perceptions of themselves and the world.

The issue of privacy also emerges prominently in the literature. Many AI systems collect and analyse sensitive mental health data, raising concerns about surveillance, data misuse and confidentiality. Barda et al. (2022) emphasise that vulnerable individuals may experience heightened anxiety when they suspect or discover that their personal information is being used in ways they do not fully understand. This anxiety can interfere with the therapeutic benefits of digital tools, reducing trust and discouraging consistent engagement. These findings suggest that psychological risks are not merely incidental but intrinsic to the operation of many algorithmic systems. Consequently, scholars emphasise the need for robust safeguards, transparent design principles and clear communication strategies to mitigate these risks.

Equity, Representation and Structural Concerns

Beyond individual risks, the literature consistently highlights broader structural challenges related to equity and representation in AI-driven mental health tools. The World Health Organization (2021) warns that AI systems used in health contexts may reinforce existing inequalities if they are trained on datasets that underrepresent marginalised communities. When algorithms interpret mental health data, their accuracy depends on the diversity of the populations included in training samples. If certain groups are missing or poorly represented, the system's ability to accurately detect and respond to their emotional states may be compromised. This problem is significant because many vulnerable populations, including racial minorities, people with disabilities, rural communities and individuals with lower socioeconomic status, are often underrepresented in digital datasets. Scholars argue that biased datasets lead to biased predictions, which can produce differential treatment outcomes and widen health disparities (Shaw et al., 2023). In mental health contexts, such biases may manifest as incorrect risk assessments, misinterpretation of tone or incomplete recognition of distress signals. Equity concerns also extend to technological access. While AI-driven mental health tools are often praised for their accessibility, not all vulnerable groups have equal opportunities to use them. Older adults or individuals with limited digital literacy may find these systems confusing or inaccessible. This digital divide may deepen existing inequalities, as those who could benefit most from additional support may be the least able to access it (Meier & Reinecke, 2021).

Lack of Focus on Differential Vulnerability

A recurring theme across the literature is the limited number of studies examining how AI affects specific vulnerable groups differently. Many studies focus on general user populations, treating vulnerability as a secondary consideration. As a result, questions about differential vulnerability remain largely unexplored. For example, the psychological risks experienced by adolescents are likely distinct from those experienced by older adults, individuals with disabilities or people facing chronic mental health conditions. Yet the literature rarely disaggregates findings across these categories (Inkster et al., 2023). This gap makes it difficult for policymakers and designers to develop targeted interventions that address the needs of specific populations. Without detailed evidence, it becomes challenging to anticipate how different users may respond to AI-driven interactions or to design safeguards that account for diverse emotional and cognitive profiles.

Algorithmic Design and Long-Term Behavioural Influence

In addition to gaps related to differential vulnerability, the literature also reveals limited examination of how algorithmic design shapes long-term behaviour. Many studies focus on short-term emotional reactions or immediate psychological outcomes, without considering how sustained engagement with AI may alter behaviour over months or years. Scholars emphasise that algorithms are not neutral; they are designed to shape user behaviour in ways that align with organisational objectives, such as increasing time spent on a platform or encouraging consistent engagement (Kelly et al., 2023). For vulnerable individuals, long-term exposure to such design features could have significant consequences. Compulsive use of AI tools, reliance on automated emotional support or repeated exposure to harmful content may gradually alter coping strategies, social habits or emotional resilience. Torous et al. (2021) argue that long-term behavioural shifts represent one of the most pressing but least understood risks in digital mental health.

THEORETICAL FRAMEWORK

This study draws on two theoretical perspectives to make sense of how AI-mediated systems affect emotionally vulnerable users: the Digital Well-Being Framework and Vulnerability Theory. Together, they provide a lens for understanding both the psychological impact of design choices and the structural inequalities that make certain people more exposed to risk.

Digital Well-Being Framework

The Digital Well-Being Framework centres on how technology design shapes users' psychological outcomes. It emphasises that interface features, feedback loops, and interaction patterns can either support or undermine mental health (Shin, 2025). For example, in AI-mediated platforms, design elements such as real-time conversational responsiveness, personalized recommendations, or adaptive micro-tasks can trigger strong emotional engagement, for better or worse (M. Peters, 2021). Recent scholarship argues that digital well-being is not just about reducing "screen time" but about enabling balanced, deliberate and healthy technology use (Discover Social Science & Health, 2025). This includes designing for user autonomy, competence, and relatedness, psychological needs that, when satisfied, contribute to wellness rather than compulsive behaviour (Peters, 2021). Interface designs that support these needs can help users feel more in control, enhance their digital literacy, and foster meaningful connections, rather than simply maximizing engagement (Peters, 2021). In AI systems specifically, digital well-being is challenged by feedback loops and algorithmic stimuli. These systems can modulate users' emotional states by adapting content and conversational tone in response to user behaviour, sometimes reinforcing negative patterns unintentionally (Adanyin, 2024). The user's internal state (e.g., emotion, motivation) responds to these stimuli, creating a relational dynamic: the more the system "knows" about the user, the more tailored (and potentially manipulative) its responses become (Shin, 2025). A human-centred AI model, informed by this framework, would treat digital well-being as a core design objective, not an afterthought.

Vulnerability Theory

Vulnerability Theory helps explain why certain users may be more susceptible to harm in AI-mediated environments. In this context, "vulnerability" refers not only to emotional or psychological fragility but also to social, cognitive, and structural dimensions that reduce a person's capacity to protect themselves (WHO, 2025). Vulnerable individuals may include adolescents, older adults, or those with limited digital literacy or mental health conditions. From this perspective, risk is not evenly distributed. Some users are structurally disadvantaged: they may lack digital self-control, or they might not understand how feedback loops influence their behaviour (AI & Society, 2024). Others may face socio-normative vulnerabilities: normative expectations about technology use, social pressure, or design manipulations that exploit cognitive biases (Technological Forecasting & Social Change, 2022). These factors deepen their dependency on AI systems and impair their ability to

disengage. Vulnerability Theory also highlights how AI systems can exacerbate preexisting inequalities. Users who are less digitally literate may not recognise addictive design features or demand safeguards. Those who are socially isolated or emotionally fragile may misinterpret AI-driven empathy as genuine human connection, increasing their risk of emotional reliance (Jiang, 2024). This suggests that design and regulation must not assume a "one-size-fits-all" user but rather account for uneven capacities, power, and agency.

Integrating the Two Frameworks

When combined, the Digital Well-Being Framework and Vulnerability Theory provide a robust foundation for analysing AI's psychological and social effects. The Digital Well-Being Framework helps us understand how design features influence emotional states, user behaviour, and long-term mental health. Vulnerability Theory clarifies why some users are more exposed to risk: their social, cognitive, or structural context reduces their resilience. Together, these theories justify the need for protective design principles and regulatory mechanisms. For instance, interfaces should be built not just to optimize engagement, but to support digital competence and autonomy so users can maintain agency (Peters, 2021; Shin, 2025). At the same time, policymakers and designers should proactively address structural vulnerabilities, for example, by ensuring transparency about feedback loops, offering user education, and creating opt-out mechanisms for users most at risk (AI & Society, 2024; WHO, 2025). Moreover, this theoretical lens underscores why human-centred AI matters. Rather than focusing exclusively on performance or efficiency, a well-being–oriented design calls for systems that encourage emotional autonomy, respect vulnerability, and build trust (Shin, 2025). Regulatory strategies informed by these perspectives could require algorithmic transparency, regular impact assessments, and inclusive participation in design from vulnerable groups.

METHODOLOGY

This study uses a qualitative document analysis approach to investigate how artificial intelligence shapes emotional wellbeing and mental health, particularly for vulnerable populations. Document analysis is a well-established qualitative research method that involves systematically reviewing, interpreting, and coding texts such as peer-reviewed research articles, policy documents, reports, and other archival materials (Morgan, 2022; Kutsyuruba, 2023). Because the analysis focuses on existing texts rather than individuals, we

did not collect personal data or interact with participants directly; therefore, ethical approval was not required.

Data Sources and Selection

To build a robust foundation of evidence, the study draws on three types of documents:

1. ***Academic literature***: Peer-reviewed research papers published between 2021 and 2024 that address AI and mental health, psychological wellbeing, or algorithmic risk.
2. ***Policy and technical documents***: Reports, white papers, government guidelines or frameworks that explicitly deal with AI in mental health or digital wellbeing during the same period.
3. ***Professional perspectives***: Articles, frameworks or commentaries produced by mental health institutions or practitioners on AI adoption in care (e.g., journal articles on professionals' views). For instance, the qualitative descriptive study by mental health professionals on AI adoption provided by Zhang et al. (2023) was included as a data source.

Documents were selected using inclusion criteria that balanced recency, relevance, and credibility: (a) published in English, (b) specifically focused on mental health, emotional wellbeing or risks associated with AI, and (c) in public domain via academic databases or official policy repositories.

Data Extraction and Analysis

Once documents were gathered, we conducted thematic coding, a process of identifying recurring patterns, concepts and concerns. Using a reflexive thematic analysis approach, we coded texts for themes such as emotional impact, privacy, content curation, algorithmic exposure, user dependency, and vulnerabilities (Morgan, 2022). Codes were developed inductively: as we reviewed more documents, new themes emerged, and earlier entries were re-examined for consistency and refinement. To ensure rigor in our analysis, we adopted procedural steps recommended in qualitative document research. This involved repeated reading of texts, memoing to capture reflexive notes, triangulation across different types of documents, and constant comparison to ensure reliability (Kutsyuruba, 2023; Chanda, 2021). We maintained a codebook to document definitions, examples, and any changes to coding as the analysis progressed.

Strengths and Limitations

One strength of this method is that it allows broad synthesis of existing literature and policy without needing new empirical data, especially useful given the rapidly evolving domain of AI in mental health. Furthermore, because no human subjects were involved, the study avoids potential ethical issues like privacy risk or participant burden (Morgan, 2022). However, document analysis also has limitations. The study depends on publicly available texts, which may introduce bias: not all institutional or proprietary AI systems publish internal reports or technical design documents. There may also be uneven geographical representation, policy documents are more accessible from certain regions, which can skew findings. In addition, interpretation of text is inherently subjective, even when coding is systematic and reflexive.

RESULTS

The document analysis revealed four major, interrelated themes around AI's mental health impact: (1) expanded access to support; (2) heightened psychological risk; (3) emotional dependency; and (4) limited transparency and user understanding. Each of these themes captures opportunities and significant challenges, particularly for users who may be emotionally or socially vulnerable.

Expanded Access to Mental Health Support

One of the clearest benefits emerging from the literature is that AI-based tools expand access to mental health support in scalable ways. AI chatbots and mobile applications can operate around the clock, offering users mood tracking, behavioural monitoring, and basic emotional guidance when traditional mental healthcare is unavailable or difficult to reach (Wasil et al., 2021). For people in remote areas, or for whom mental health services are prohibitively expensive or stigmatized, the sheer availability of these tools can make a critical difference. Research supports that these tools are not just accessible but also effective. In a systematic review and meta-analysis of 31 randomized controlled trials involving nearly 30,000 adolescents and young adults, AI chatbots demonstrated small-to-moderate improvements in depression, anxiety, stress, and psychosomatic symptoms (PubMed, 2024). These results suggest that, at least in the short term, AI mental health tools can deliver meaningful symptom relief for a broad user base (PubMed, 2024). Beyond symptom management, newer generative AI models show promise in building a therapeutic alliance akin to human support. A recent cohort study of a generative AI designed specifically for mental health found significant reductions in self-reported depression (PHQ-9) and anxiety (GAD-7) over a 10-

week period. The study also reported improvements in social interaction, hope, perceived social support, and decreased loneliness, suggesting that these tools can provide not only emotional intervention but also a sense of social connectedness (Hull et al., 2025). Moreover, AI can help public health systems operate more effectively. By aggregating anonymized data from user interactions, platforms can identify trends in mental health, flag emerging issues, and direct resources proactively. Such predictive analytics are particularly useful in low-resource settings where mental health professionals are scarce; AI can act as a frontline detection tool, helping systems scale preventive care (Barda et al., 2022). These findings demonstrate that AI has substantial potential as a supplement, not a substitute, for traditional mental health care. For many users, especially those marginalized in existing healthcare systems, AI can offer a lifeline of support, early detection, and ongoing companionship.

Heightened Exposure to Psychological Risk

Despite its benefits, the literature reveals significant psychological risks associated with AI-mediated mental health support. One major concern: vulnerable users are more likely to be exposed to harmful or emotionally destabilizing content through algorithmic curation. Algorithms designed to maximize engagement may disproportionately surface emotionally intense or negative content. In AI-driven social media contexts, this curation can reinforce insecurity, self-doubt, and compulsive behaviours. Research into Generation Z's experience with algorithmic content shows that emotion-triggering negative content is often prioritized, contributing to a "loop" that deepens emotional instability (Nguyen et al., 2024). For example, the authors of a recent MDPI review found that AI systems may amplify content that evokes fear or distress, which in turn worsens mental health outcomes among susceptible users (Nguyen et al., 2024). Adolescents are particularly vulnerable to these algorithmic harms. A recent letter published in the *Asian Journal of Psychiatry* warns that AI-driven social media may exacerbate anxiety, depression, self-esteem issues, and body dissatisfaction in teenagers. The authors call for more research into how engagement-prediction tools and real-time behaviour analysis shape adolescent mental health (Asian Journal of Psychiatry, 2025). This concern is echoed by broader interdisciplinary work, which links social media algorithms to addiction and adverse mental health outcomes among youth (American Journal of Law & Medicine, 2023). Empirical studies corroborate these theoretical risks. A 2024 study of school-aged children found that exposure to a variety of social media threats, including harassment, misinformation and "appearance pressure" content, was strongly associated with depressive symptoms and anxiety (Child & Adolescent Psychiatry & Mental

Health, 2024). Although not all these threats derive from AI chatbots per se, they illustrate the broader ecosystem in which algorithmic curation interacts with youth vulnerabilities. When AI chatbots are situated within or alongside these algorithmic systems, the emotional risk intensifies. Ethical analyses highlight that conversational agents may reinforce negative thought patterns or deepen emotional distress, especially if they lack nuance or deliver unsafe advice (JMIR Mental Health, 2025). These ethical challenges are compounded by the fact that only a small minority of studies empirically examine the perspectives of users with mental health conditions, meaning safety concerns often go under documented (JMIR Mental Health, 2025).

Emotional Dependency on AI Systems

Perhaps one of the most profound and alarming findings relates to emotional dependency. Vulnerable users sometimes begin to rely heavily on AI chatbots not only for emotional support, but for companionship and validation in ways that mirror human relationships, and that can displace real interpersonal connection. In a recent longitudinal randomized controlled study, researchers examined how different modes of chatbot interaction (text, neutral voice, engaging voice) and content type (personal, open-ended, non-personal) affected psychosocial outcomes over four weeks (Fang et al., 2025). While voice-based chatbots initially reduced loneliness more than text alone, high-frequency usage ultimately correlated with greater emotional dependence, decreased socialization, and increased problematic use. Users who started with higher baseline attachment tendencies or trust in the AI experienced sharper increases in dependency over time (Fang et al., 2025). Complementing this, theoretical work on “technological folie à deux” highlights how certain users, particularly those with mental health vulnerabilities, may experience destabilized belief systems when engaging deeply with chatbots (Dohnány et al., 2025). The authors describe feedback loops in which a user’s mental illness amplifies an AI’s agreeableness, which in turn can reinforce delusional or distorted thinking. Over time, this dynamic may undermine the user’s ability to reality-test or maintain psychological boundaries (Dohnány et al., 2025). Empirical and normative research also documents relational risks: AI companionship may lead to idealized attachment, overestimation of the AI's understanding and underestimation of the risks. In a qualitative and design-focused study, Ngwenyama et al. (2024) show that anthropomorphic chatbots often draw users into a Faustian bargain, users trade autonomy and emotional self-governance for constant engagement and connection with a non-human entity. These relationships can disrupt real-life social ties and emotional regulation (Ngwenyama et al., 2024). Meanwhile,

clinicians and policy experts raise red flags. Psychotherapists have reported seeing clients replace or deprioritize human contact in favour of AI-based "companions," leading to isolation, increased distress and cognitive distortion (Guardian, 2025). As one expert put it, the loss of "safe space", where a person feels truly heard, is a serious concern if therapy becomes dominated by algorithmic voices (Guardian, 2025).

Limited Transparency and User Understanding

A final major finding concerns the opacity of AI systems and the limited understanding many users have about how their data is used and how decisions are made. This lack of transparency produces anxiety, distrust, and a sense of lost control, especially among more vulnerable populations.

First, many users are unaware of how conversational agents collect, process, and respond to their data. Scoping reviews of the literature highlight that privacy and confidentiality are among the most common ethical concerns in AI-mediated mental health (JMIR Mental Health, 2025). Users may not realize that chatbots log conversation data, track emotional states, or feed back into broader model training systems. This black-box nature is deeply problematic when the content is sensitive and personal.

Second, chatbots sometimes cannot clearly explain how they generate responses. Because they operate via large models trained on vast, heterogeneous data, the reasoning behind their suggestions is often opaque, even to researchers. Without a clear rationale, users may question whether the advice is trustworthy. This doubt can fuel anxiety and mistrust, undermining the therapeutic benefit of AI (Psychology Today, 2025).

Third, the regulatory and ethical infrastructure of these tools lags behind their technical capabilities. While some developers build guardrails, others do not, and many users are never informed of the limitations. This inconsistency raises serious accountability issues. For example, a recent review underscores that conversational AI lacks standardized mechanisms for crisis detection, escalation, or psychiatric referral (JMIR Mental Health, 2025).

Fourth, design choices in AI interfaces often exploit vulnerability without clear user consent or comprehension. Ngwenyama et al. (2024) argue that emotionally engaging chatbots can manipulate relational cues to foster deeper attachment, all while obscuring their transactional nature. Users may anthropomorphize these tools and form attachments, but they often remain

unaware that the system's "empathy" is an engineered simulation, not true human care (Ngwenyama et al., 2024).

Fifth, a few emerging technological solutions aim to improve transparency, but they are not yet widespread. For instance, a recent proposal called EmoAgent introduces a multi-agent AI system that simulates vulnerable user interactions to detect risk and intercede when necessary. EmoAgent's design includes components to monitor, predict, and provide corrective feedback when a user's mental state deteriorates (Qiu et al., 2025). While innovative, such safeguards remain mostly in the research phase and are not yet broadly adopted.

Finally, this lack of transparency contributes to a broader sense of loss of control. When people do not know how their data is stored, how decisions about emotional content are made, or who can access their conversation history, they may feel disempowered. This, in turn, can erode trust in AI systems and damage users' relationship to their own emotional autonomy (JMIR Mental Health, 2025; Guardian, 2025).

Synthesis: Interplay Among Themes

When considered together, these four themes reveal a deeply ambivalent impact of AI on mental health, especially for emotionally vulnerable users. On one hand, AI tools offer unprecedented access to support and can deliver clinically meaningful benefits in symptom reduction. On the other, they expose people to new forms of risk that stem from design, dependency and opacity. The expanded access that AI affords is powerful but should not obscure its limitations. For many users, the tools function as a stopgap rather than a substitute for professional care, and they work best when embedded within a larger support ecosystem. Meanwhile, algorithmic curation and emotional dependency highlight how AI systems can transform vulnerable users' internal and social worlds, sometimes in ways that deepen rather than alleviate distress. Transparency issues, and the resulting lack of user understanding, complicate these dynamics further. Users may not know whether they are interacting with a benign companion or a poorly regulated system with hidden risks. Without clear frameworks for accountability and user protection, even the best-intentioned tools may inflict psychological harm. These findings suggest the need for more nuanced design, stronger regulation, and greater education for users. Emotional safety must become a core priority for designers and policymakers: AI tools should not only be effective and scalable, but also

transparent, accountable, socially aware and attuned to the complexity of human vulnerability.

DISCUSSION

The results of this review paint a complex, ambivalent picture: artificial intelligence offers real promise to support vulnerable users, but it also introduces serious risks, many of which hit the most emotionally or socially fragile people hardest. In interpreting these findings, three interlocking considerations stand out. First, the advantages of access and early detection are powerful, but they must be balanced carefully against the potential harms arising from opaque systems, exploitative design, and weak regulatory guardrails. Second, the most vulnerable users, those with limited digital literacy, social isolation, or pre-existing mental health challenges, often lack the resources to navigate these systems safely. Third, designers, regulators, and mental health practitioners need to take collective responsibility for creating protective frameworks, not just technical fixes.

Balancing Benefits and Harms

One of the most promising aspects of AI in mental health is the way it expands accessibility. The review highlighted that conversational agents and apps can offer 24/7 support, mood tracking, and behavioural monitoring, filling gaps in traditional mental health systems. These capabilities are especially valuable in settings where mental health professionals are scarce or where stigma and cost prevent people from seeking help (Wang et al., 2025). The ability of AI to provide early detection, by flagging shifts in mood or pattern of use, is also a potentially transformative safeguard, particularly for those who might otherwise go unnoticed by services. Yet, this benefit comes with significant downside risks. The review found that, for some users, the same systems that offer availability and engagement can also deepen emotional vulnerability. AI designs that optimize for engagement may surface emotionally provocative content or reinforcing patterns, amplifying loneliness or compulsive use. The persuasive nature of algorithmically curated advice, especially when combined with conversational AI, may exploit emotional needs in ways that increase risk rather than reduce it. This trade-off, access versus risk, is not simply technical. It is deeply ethical, and it demands that designers, clinicians, and policymakers think carefully about how to create systems that help rather than harm. Accepting accessibility gains without addressing how systems might be misused or misinterpreted by vulnerable individuals risks exposing people to new forms of psychological harm.

Vulnerability: Beyond Individual Factors

The findings emphasize that the people most likely to experience harm from AI mental health tools are often those least equipped to protect themselves. Vulnerability does not stem solely from clinical diagnoses; it also arises from structural, social, and cognitive conditions. Research on social anxiety, for example, shows that users who struggle with loneliness and rumination may develop problematic use of conversational AI. A recent study found that social anxiety was positively associated with "Problematic Use of Conversational AI" (PUCAI), and that this relationship was mediated by loneliness and rumination (Kwon et al., 2023). Users who perceive mind and intention in their AI interlocutors ("mind perception") were especially at risk, because their emotional investment is deeper, making them more susceptible to over-reliance. On another front, the concept of a "technological folie à deux" has been introduced to describe worrying feedback loops between AI chatbots and users with mental illness (Dohnány et al., 2025). In such dynamics, a user's cognitive vulnerabilities, such as impaired reality testing or skewed belief updating, interact with a chatbot's adaptive agreeableness. Over time, this can destabilize belief systems, erode psychological resilience, and amplify delusional thinking or emotional dependency. These insights underscore that regulatory measures or design interventions that assume a "typical" user will not suffice. Instead, we need approaches that acknowledge and respond to differential vulnerability: some people will form deep emotional attachments, others may over-disclose, and still others may lack the literacy to question or control how their data is used.

The Role of Transparency, Education, and Regulation

Because these risks are not purely individual, they call for systemic solutions. AI designers and mental health policymakers must implement protective measures that address both technical and social dimensions.

Transparent Data Practices

Many users are unaware of what data is collected, how it is processed, or who has access to their conversations. The scoping review of AI ethics found that poor transparency contributes to mistrust and can exacerbate anxiety (Meadi et al., 2025). Without clear disclosures or understandable privacy controls, users may overshare or be manipulated by systems that are not designed to prioritize their well-being. To counteract this, developers should adopt more transparent architectures: explaining in plain language what data is stored, how it's used, and

how users can control or delete it. This is not only a technical design issue, but a matter of informed consent and user autonomy.

User Education Initiatives

Vulnerable users often do not have the digital literacy to understand how AI systems operate or how to self-regulate their use. This gap points to the need for embedded AI-literacy frameworks: systems that teach users about data risks, over-disclosure, and emotional boundaries within AI interactions (Anvari & Wehbe, 2025). By embedding these principles into AI tools themselves, through guided dialogues, teaching modules, or onboarding flows, developers can empower users to use systems more safely.

Inclusive and Ethical Design

Designers should prioritize inclusive design that prevents harm for high-risk users. For instance, AI chatbots could incorporate different modes that limit emotional intensity or limit usage for users prone to overuse. The randomized trial by Fang et al. (2025) shows that voice-based chatbots may initially reduce loneliness better than text, but when usage becomes excessive, they increase emotional dependence and problematic use. Designers should build safeguards into AI based on such evidence: limiting frequency, providing reminders, or enabling "cool-down" periods.

Regulatory and Safety Frameworks

Policymakers have a critical role in establishing guardrails. Regulatory challenges already exist: for example, in South Africa, mental health apps collect sensitive user data (behavioural patterns, emotional states), but legal frameworks do not mandate robust safeguarding (Frontiers, 2025). Without enforceable regulations around data de-identification, encryption, and third-party sharing, vulnerable users remain exposed to exploitation or data misuse. Regulators should also require safety protocols in digital mental health tools: for instance, mandating crisis-detection capabilities, escalation mechanisms to human support, or independent audit requirements for large-scale AI chatbots. Ethical design standards could mandate transparency reports and AI-literacy education mechanisms embedded in the tools themselves (Shehab, 2025).

Theoretical Implications and Broader Social Context

The theoretical lenses of this study, Digital Well-Being and Vulnerability Theory, offer compelling insight into why these design and regulatory issues matter. The Digital Well-Being perspective helps us understand how system architecture influences emotional states and behavioural patterns. When AI tools are designed to maximize engagement or emotional resonance, they may inadvertently exploit users' psychological needs, amplifying loneliness,

reinforcing negative beliefs, or prompting over-disclosure. Vulnerability Theory complements this by reminding us that risk is not distributed evenly. Some people are more susceptible because of social isolation, cognitive differences, or lack of literacy. Their inability to negotiate emotional boundaries with AI systems increases their exposure to harm. Together, these theories suggest that effective governance cannot rely solely on technical fixes; it must also address structural and social inequalities. Regulators and designers must account for the uneven distribution of risk, ensuring that systems are built to protect and support, rather than merely engage or scale.

Research Gaps and Future Directions

The findings point to several critical gaps in the existing scholarship, and urgent directions for future work.

Empirical Studies on Vulnerable Populations

There is a clear lack of long-term, empirical research that centers on the most vulnerable groups: adolescents with mental health challenges, older adults, socioeconomically disadvantaged individuals, or people with low digital literacy. Most existing randomized controlled trials (RCTs) measure short-term symptom reduction (e.g., anxiety, depression), but few explore how dependency or relational risks develop over months or years. Addressing this gap will require longitudinal studies, mixed-methods research, and psychosocial evaluations tailored to high-risk populations.

Dependency Trajectories

Emotional reliance on AI companionship is a novel phenomenon that poses complex developmental risks. How do dependency behaviours form and evolve? What personality traits or external conditions predict problematic use? Research such as Fang et al. (2025) provides an important first step, but more nuanced investigations are needed, studies that assess when and how users' relationships with AI transition from therapeutic to potentially harmful.

Ethics and Safety Protocols

There is a need for more evaluation of ethical and safety protocols. Which guardrails work best? Should regulators require that all mental health chatbots include crisis escalation, human fallback, or usage caps? Comparative studies of different safety architectures (e.g., age gating, voice modulation, usage limits) could inform best practices.

AI Literacy Interventions

Because over-disclosure and misunderstanding of data practices are already documented risks, research should test interventions that embed AI literacy inside the user experience. Do guided disclosures, interactive consent modules, or in-app tutorials improve user autonomy and reduce harm? Anvari and Wehbe (2025) have proposed an embedded AI-literacy framework; empirical testing of such models will be essential.

Policy and Governance Innovation

Finally, regulatory innovation is needed to keep pace with technology. Scholars and policymakers should collaborate to design adaptive governance models that include transparency requirements, data protection standards, and independent oversight for mental health AI. Research should track the implementation and outcomes of regulatory experiments to determine what works and what can be scaled in different jurisdictions.

Ethical and Social Implications

The ethical stakes of this topic are high. For vulnerable users, AI chatbots are not just tools, they can become companion-like entities, forming emotional ties. This raises deep questions about agency, autonomy, and what it means to be human in a world where machines mirror our emotional expressions. If poorly regulated, AI could exploit loneliness, data vulnerabilities, or emotional fragility, turning technology into a subtle mechanism of control rather than care. At the same time, abandoning AI as a mental health tool would be equally irresponsible. These technologies offer unprecedented scale and reach, particularly in low-resource settings. The goal should not be to reject AI, but to shape it: to develop emotionally intelligent systems that respect user dignity, promote mental wellbeing, and remain transparent and accountable.

CONCLUSION OF THE DISCUSSION

In sum, this study underscores that AI's role in mental health is deeply ambivalent. While its capacity to deliver scalable, accessible, and immediate emotional support is a powerful promise, especially for underserved or vulnerable groups, there are serious psychological and structural risks that cannot be ignored. Emotional dependency, algorithmic opacity, and exploitative data practices are not mere side-effects; they are consequences of design and regulation choices. Moving forward, a multi-pronged approach is required. Designers must build with empathy and constraint; users must be empowered with literacy; regulators must demand transparency, safety, and equity; and researchers must study how dependency and harm evolve over time. Importantly, interventions must not be one-size-fits-all: real

protection requires recognizing the diversity of vulnerability and building AI systems that uplift without undermining the emotional lives of the people they serve.

CONCLUSION

Artificial intelligence has become deeply woven into the everyday experiences of people around the world. It shapes how individuals communicate, search for information, manage their emotions and seek support during periods of distress. As AI tools move rapidly into health and wellbeing spaces, they are increasingly influencing how people understand their mental states and how they cope with psychological strain. This study set out to examine these shifts, with particular focus on the ways vulnerable users navigate the growing ecosystem of AI-driven mental health tools. What emerges is a nuanced picture of both promise and peril. The expansion of AI-based mental health support represents an important development in a world where access to psychological care remains unequal. Many communities face severe shortages of mental health professionals, long waiting times or financial barriers that make traditional therapy difficult. Within this context, AI chatbots, mobile applications and virtual assistants offer alternatives that can provide immediate and cost-effective guidance. They create opportunities for individuals who might otherwise remain unsupported. Users can monitor their mood, track behavioural patterns or express emotions at any time of day, without the fear of stigma or judgement. For individuals who are socially isolated, overstretched, or living in under-resourced environments, these tools can be a meaningful source of comfort and connection.

However, the benefits of accessibility must be considered alongside substantial risks. The review shows that AI systems do not affect all users in the same way. Vulnerable users, especially those facing loneliness, chronic stress, low digital literacy or pre-existing psychological challenges, appear more likely to experience harm. One of the most concerning risks is emotional manipulation. Many AI systems rely on design features that aim to increase engagement. These systems respond empathetically, adapt to a user’s tone, and sustain conversations that feel personalised. While this can make interactions feel supportive, it can also blur emotional boundaries. Some users may interpret the system's tone as genuine care and gradually form attachments that overshadow human relationships. As dependence deepens, the distinction between technological support and emotional reliance becomes increasingly difficult to maintain. Exposure to harmful or emotionally charged content is another challenge. Algorithms curate information based on patterns of interaction, and these

patterns may expose vulnerable users to content that reinforces negative emotions, anxieties or compulsive behaviours. A user who expresses sadness may be shown more material that mirrors or deepens that emotion. The result is an algorithmically constructed feedback loop where a person's vulnerabilities are unintentionally amplified rather than alleviated. This dynamic can produce real psychological strain and may prolong emotional distress rather than assist with recovery.

Dependency on AI systems represents a third significant concern. While many users turn to AI tools for short-term relief, some begin using them as primary sources of emotional support. Over time, this can reduce motivation to seek human connection or professional help. Relationships with AI systems may feel predictable and safe, but they risk replacing the complexity and reciprocity that characterise healthy human interactions. Dependence can also undermine long-term wellbeing by weakening coping strategies, reducing resilience and lowering the threshold for turning to technological solutions rather than developing interpersonal support networks. Underlying all these risks is the issue of opacity. Most people do not fully understand how AI systems make decisions or what happens to the information they share. This lack of clarity extends from data collection to algorithmic processing and storage. Vulnerable individuals may disclose highly sensitive details about their emotional life without knowing how secure the data is or how it may be used in the future. Not understanding these processes can contribute to mistrust, anxiety and a sense of losing control over one's personal information. Ethical concerns become even more pronounced when considering the possibility of data sharing with third parties, commercial exploitation or profiling based on psychological patterns.

The findings of this study contribute to discussions on digital wellbeing by highlighting the importance of recognising and addressing these vulnerabilities. Digital wellbeing is not only about reducing screen time or managing notifications. It is also concerned with ensuring that technologies support a person's emotional, cognitive and relational stability. For AI systems used in mental health contexts, wellbeing must be understood as a multidimensional outcome that depends on system design, user literacy, regulatory frameworks and broader social environments. One key contribution of this study is its emphasis on the need for design interventions that protect vulnerable users. AI systems should be built with safeguards that prevent overreliance, limit emotional intensity and provide clear explanations of how data is used. Developers should prioritise interfaces that encourage reflection, rather than impulsive

engagement. Features such as usage reminders, simplified data summaries and emotionally neutral responses can reduce risks without undermining the utility of the systems. Importantly, design should be inclusive, accounting for diverse user needs and levels of digital literacy. Tools must be accessible not only in the technical sense, but also in ways that promote safe and informed use.

Policy interventions are equally essential. Regulatory frameworks need to address data protection, transparency and accountability for AI systems involved in mental health support. This includes clear guidelines on data retention, criteria for consent, and responsibilities for managing emotional risk. Policymakers should also establish standards for evaluating the safety and effectiveness of AI-based mental health tools before they reach the public. As AI technologies continue to evolve, regulatory mechanisms must remain adaptable and informed by ongoing research, including studies that explore psychological, social and ethical impacts. Looking forward, this study identifies several important directions for future research. One priority is the need for longitudinal studies that trace how user experiences unfold over months or years. Short-term evaluations may overlook slow-developing patterns of dependency or emotional attachment. Long-term data would allow researchers to understand how AI influences mental health trajectories, coping mechanisms and social relationships over extended periods.

Another critical area is participatory design, where users, especially vulnerable groups, are meaningfully involved in shaping AI tools. Their lived experiences can guide developers toward features that support rather than undermine wellbeing. Participatory approaches also strengthen ethical integrity by ensuring that systems are aligned with real-world needs rather than assumptions made from a technical perspective. Future research must also explore how vulnerability intersects with age, gender, socioeconomic status, digital literacy and psychological health. Vulnerability is not a single characteristic, but a layered and evolving condition shaped by personal and structural factors. Children, older adults, people living with chronic mental health conditions, and those facing economic hardship may each encounter AI systems differently. Understanding these variations is crucial for designing equitable technologies that do not worsen existing inequalities.

In conclusion, artificial intelligence is reshaping the landscape of mental health support in profound ways. It offers opportunities to broaden access, reduce stigma and provide timely emotional assistance. At the same time, it brings a set of risks that disproportionately affect

those who are already vulnerable. This study demonstrates that safeguarding mental wellbeing in the age of AI requires coordinated effort: thoughtful design, robust policy, and research that keeps pace with rapid technological change. By centering the experiences of vulnerable users, society can work toward AI systems that genuinely support psychological wellbeing while preserving autonomy, dignity and human connection.

REFERENCES

1. Adanyin, A. (2024). *AI-driven feedback loops in digital technologies: Psychological impacts on user behaviour and well-being*. arXiv. https://doi.org/10.48550/arXiv.2411.09706
2. American Journal of Law & Medicine. (2023). *Algorithms, addiction, and adolescent mental health: An interdisciplinary study to inform state-level policy action to protect youth from the dangers of social media*. *American Journal of Law & Medicine, 49*(2–3), 135–172. https://doi.org/10.1017/amj.2023.25
3. Anvari, S. S., & Wehbe, R. R. (2025). *Therapeutic AI and the hidden risks of over-disclosure: An embedded AI-literacy framework for mental health privacy*. arXiv. https://doi.org/10.48550/arXiv.2510.10805
4. Asian Journal of Psychiatry. (2025). *The need for research on AI-driven social media and adolescent mental health*. *Asian Journal of Psychiatry, 108*, Article 104513. https://doi.org/10.1016/j.ajp.2025.104513
5. Barda, A., Sela, T., Elisha, D., & Schuster, A. (2022). AI privacy risks in personalised digital services. *AI and Society, 37*(3), 987–1001. https://doi.org/10.1007/s00146-021-01235-5
6. Chanda, A. (2021). *Key methods used in qualitative document analysis*. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3996213
7. Child & Adolescent Psychiatry & Mental Health. (2024). *Social media threats and health among adolescents: Evidence from the Health Behaviour in School-aged Children Study*. *Child & Adolescent Psychiatry & Mental Health, 18*, Article 62. https://doi.org/10.1186/s13034-024-00754-8
8. Discover Social Science & Health. (2025). *Understanding digital wellbeing: Impacts, strategies, and the path to healthier technology practices*. *Discover Social Science & Health, 5*, Article 145. https://doi.org/10.1007/s44155-025-00259-5
9. Dohnány, 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. https://doi.org/10.48550/arXiv.2507.19218

10. Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). *How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study*. arXiv. https://doi.org/10.48550/arXiv.2503.17473
11. Fleming, T., Bavin, L., & Lucassen, M. (2022). Adolescent experiences of AI-based counselling systems. *Journal of Adolescent Health, 71*(3), 456–464. https://doi.org/10.1016/j.jadohealth.2022.04.012
12. Frontiers in Pharmacology. (2025). *Regulatory challenges of digital health: The case of mental health applications and personal data in South Africa*. *Frontiers in Pharmacology, 16*, Article 1498600. https://doi.org/10.3389/fphar.2025.1498600
13. Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research, 34*(7), 595–606. https://doi.org/10.1177/10497323231217392
14. Hull, T. D., Zhang, L., Arean, P. A., & Malgaroli, M. (2025). *Mental health generative AI is safe, promotes social health, and reduces depression and anxiety: Real world evidence from a naturalistic cohort*. arXiv. https://doi.org/10.48550/arXiv.2511.11689
15. Inkster, B., Evans, J., & Finkelstein, J. (2023). The effectiveness of AI-based mental health tools: A systematic review. *Lancet Digital Health, 5*(4), e234–e245. https://doi.org/10.1016/S2589-7500(23)00012-6
16. Jiang, Z. Z. (2024). *Self-disclosure to AI: The paradox of trust and vulnerability in human-machine interactions*. arXiv. https://doi.org/10.48550/arXiv.2412.20564
17. JMIR Mental Health. (2025). *Exploring the ethical challenges of conversational AI in mental health care: Scoping review*. *JMIR Mental Health, 12*(1), Article e60432. https://doi.org/10.2196/60432
18. Kelly, J., Martin, P., & Choi, J. (2023). Everyday AI adoption and its psychological implications. *Computers in Human Behavior, 139*, 107550. https://doi.org/10.1016/j.chb.2022.107550
19. Kutsyuruba, B. (2023). Document analysis. In *Varieties of Qualitative Research Methods* (pp. 139–146). Springer. https://doi.org/10.1007/978-3-031-04394-9_23
20. Kwon, O. H., Kim, S. Y., Lee, H., & Kim, J. (2023). How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind

perception. *Computers in Human Behavior, 145*, 107760. https://doi.org/10.1016/j.chb.2023.107760

21. MDPI. (2024). *Navigating the digital maze: A review of AI bias, social media, and mental health in Generation Z*. *MDPI, 6*(6), Article 118. https://doi.org/10.3390/2673-2688-6-118
22. MDPI. (n.d.). *AI chatbots in digital mental health*. *Computers, 10*(4), Article 82. https://doi.org/10.3390/computers10040082
23. Meadi, M. R., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the ethical challenges of conversational AI in mental health care: Scoping review. *JMIR Mental Health, 12*(1), e60432. https://doi.org/10.2196/60432
24. Meier, A., & Reinecke, L. (2021). Computer-mediated communication and mental health: A review. *Annals of the International Communication Association, 45*(2), 145–160. https://doi.org/10.1080/23808985.2021.1912096
25. Morgan, H. (2022). Conducting a qualitative document analysis. *The Qualitative Report, 27*(1). https://doi.org/10.46743/2160-3715/2022.5044
26. Ngwenyama, O. K., Ogbomo, J. O., & Boonstra, A. (2024). AI companionship or digital entrapment? Investigating the impact of anthropomorphic AI-based chatbots. *Journal of Innovation & Knowledge*. https://doi.org/10.1016/j.jik.2024.100234
27. Peters, D. (2021). *Well-being supportive design: Research-based guidelines for supporting psychological wellbeing in user experience*. Imperial College London & University of Cambridge.
28. Qiu, J., He, Y., Juan, X., Wang, Y., Liu, Y., Yao, Z., Wu, Y., Jiang, X., Yang, L., & Wang, M. (2025). *EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety*. arXiv. https://doi.org/10.48550/arXiv.2504.09689
29. Shaw, H., Ellis, D., & Ziegler, F. (2023). Technology use, vulnerability and emotional wellbeing in digital contexts. *Journal of Affective Disorders, 338*, 135–144. https://doi.org/10.1016/j.jad.2023.05.012
30. Shehab, A. (2025, March 2). *When your therapist is an algorithm: Risks of AI counseling*. Psychology Today South Africa. https://www.psychologytoday.com/za/blog/the-human-algorithm/202503/when-your-therapist-is-an-algorithm-risks-of-ai-counseling
31. Shin, Y. (2025). Toward human-centered artificial intelligence for users' digital well-being: Systematic review, synthesis, and future directions. *JMIR Human Factors, 12*, e69533. https://doi.org/10.2196/69533

32. Singh, S. H., Jiang, K., Bhasin, K., Sabharwal, A., Moukaddam, N., & Patel, A. B. (2024). *RACER: An LLM-powered methodology for scalable analysis of semi-structured mental health interviews*. arXiv. https://doi.org/10.48550/arXiv.2402.02656
33. The Guardian. (2025, August 30). 'Sliding into an abyss': Experts warn over rising use of AI for mental health support. *The Guardian*. https://www.theguardian.com/society/2025/aug/30/therapists-warn-ai-chatbots-mental-health-support
34. Torous, J., Henson, P., & Wisniewski, H. (2021). Digital mental health and user awareness of algorithmic influence. *NPJ Digital Medicine, 4*(1), 1–5. https://doi.org/10.1038/s41746-021-00460-8
35. Vaswani, P., Prabhu, J., & Lim, A. (2023). Emotional risks in AI-mediated interactions: Implications for vulnerable users. *AI Ethics, 4*(2), 75–89. https://doi.org/10.1007/s43681-022-00291-w
36. Wang, Y., Li, X., Zhang, Q., Yeung, D., & Wu, Y. (2025). Effect of a CBT-based AI chatbot on depression and loneliness in Chinese university students: Randomized controlled trial. *JMIR mHealth and uHealth, 13*, e63806. https://doi.org/10.2196/63806
37. Wasil, A., Venturo-Conerly, K., Shingleton, R., & Weisz, J. (2021). Benefits of digital mental health interventions in youth populations. *Journal of Child Psychology and Psychiatry, 62*(14), 1712–1722. https://doi.org/10.1111/jcpp.13415
38. WHO Guideline Committee. (2024). Use of qualitative research in World Health Organization guidelines: A document analysis. *Health Research Policy and Systems, 22*, 127. https://doi.org/10.1186/s12961-024-01120-y
39. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*. WHO Press.
40. World Health Organization. (2025). *Addressing the digital: Young people's technology use and mental health outcomes (Evidence review)*. WHO.
41. Zhang, M., Scandiffio, J., Younus, S., Jeyakumar, T., Karsan, I., Charow, R., … & Wiljer, D. (2023). The adoption of AI in mental health care: Perspectives from mental health professionals. *JMIR Formative Research, 7*, e47847. https://doi.org/10.2196/47847