Artificial Intelligence as a Psychological Companion: Opportunities, Challenges, and Future Directions

Sofia Elias[1]*, Nurul Aisyah Kamrozzaman[2]

Faculty of Education and Humanities, UNITAR International University, Malaysia

***Corresponding Author**

DOI : https://doi.org/10.47772/IJRISS.2026.102400013

Received: 20 August 2026; Accepted: 25 August 2026; Published: 25 September 2026

ABSTRACT

Conversational artificial intelligence (AI) has produced a new class of technology: the AI psychological companion, a system with which users maintain ongoing, emotionally significant interactions for support, comfort, and self-reflection. Millions of people now confide in companion applications such as Replika, in therapeutic chatbots such as Woebot and Wysa, and in general-purpose large language models (LLMs), a trend accelerated by a global mental health treatment gap in which most people with mental disorders receive no care. This article presents an integrative narrative review of interdisciplinary evidence on AI psychological companionship, prioritizing studies published between 2021 and 2026 and drawn from Scopus and equivalent databases. The review traces the evolution of conversational agents from rule-based scripts to generative LLMs and synthesizes the theoretical foundations of human-AI relationships, including the Computers Are Social Actors paradigm, attachment theory, and social penetration theory. The results identify five major opportunities: accessibility and scalability, reduced stigma and enhanced self-disclosure, emerging randomized controlled trial evidence of clinical benefit, loneliness mitigation, and augmentation of human care. Six challenges are then examined: emotional dependence, social displacement, unsafe crisis responses, privacy and data governance risks, algorithmic bias, and unresolved questions about artificial empathy. A conceptual framework linking system, user, and contextual factors to relational processes and divergent well-being outcomes is proposed. The discussion situates the field within United Nations Sustainable Development Goal 3 (Good Health and Well-being), particularly Target 3.4 on mental health promotion, and outlines an agenda for research, design, and governance centered on hybrid human-AI stepped-care models. The article concludes that AI companions are consequential relational technologies whose net effect on public mental health will depend on deliberate scientific, ethical, and regulatory choices.

Keywords: Artificial Intelligence, Psychological Companion, Mental Health Chatbot, Large Language Models, Sustainable Development Goal 3 (Good Health And Well-Being)

INTRODUCTION

Mental ill-health is one of the defining public health challenges of the twenty-first century. The World Health Organization (WHO) estimates that nearly one billion people live with a mental disorder, that depression and anxiety cost the global economy approximately one trillion US dollars each year in lost productivity, and that in low- and middle-income countries more than 75% of people with mental disorders receive no treatment at all (World Health Organization, 2022). Even in high-income settings, shortages of psychiatrists, psychologists, and counselors, long waiting lists, high costs, and persistent stigma leave a large share of psychological need unmet (Torous et al., 2021; Pham et al., 2022). Concerns about psychological well-being extend to higher education: recent Malaysian evidence indicates that university students face significant challenges in sustaining meaning in life and psychological wellness, underscoring the local relevance of scalable support mechanisms (Elias et al., 2025). It is against this backdrop of scarcity that a new class of technology has emerged and diffused with remarkable speed: conversational artificial intelligence (AI) systems that people use not merely as tools, but as companions.

The phenomenon is no longer marginal. Companion applications such as Replika report tens of millions of registered users; therapeutic chatbots such as Woebot and Wysa have been downloaded millions of times and evaluated in peer-reviewed trials (Fitzpatrick et al., 2017; Inkster et al., 2018); and general-purpose large language models (LLMs) such as ChatGPT are routinely repurposed by users for emotional support, venting, and quasi-therapeutic conversation (Lawrence et al., 2024; Stade et al., 2024). Students in particular have integrated AI systems into their daily academic and personal lives, even as they report difficulties and ambivalence in doing so (Jie & Kamrozzaman, 2024). Survey and interview research demonstrates that many users describe these systems as friends, confidants, and in some cases romantic partners, and report that the relationships feel subjectively real and emotionally consequential (Skjuve et al., 2021; Brandtzaeg et al., 2022; Pentina et al., 2023). In a survey of 1,006 student users of Replika, Maples et al. (2024) found that although respondents were lonelier than typical student populations, they perceived high levels of social support from the agent, and 3% spontaneously reported that the application had halted their suicidal ideation. At the same time, qualitative and experimental evidence documents real harms, including emotional dependence, distress following abrupt product changes, and unsafe responses to users in crisis (Laestadius et al., 2024; De Freitas et al., 2024).

The scholarly response to this phenomenon has been vigorous but fragmented. Human-computer interaction researchers study relationship formation with social chatbots; clinical scientists run randomized controlled trials (RCTs) of therapeutic agents; communication scholars examine disclosure and perceived empathy; ethicists debate deception, dependence, and accountability; and information systems researchers analyze adoption, engagement, and addiction (Chaturvedi et al., 2023). What is comparatively rare is integration: a synthesis that treats the AI psychological companion as a coherent object of study spanning consumer companion apps, clinical chatbots, and general-purpose LLMs, and that weighs opportunities and risks within a single conceptual frame.

This article addresses that gap. Its objectives are fourfold: (1) to define the AI psychological companion and trace its technological and theoretical lineage; (2) to synthesize recent empirical evidence, prioritizing work published between 2021 and 2026, on the psychological opportunities these systems afford; (3) to critically examine the psychological, ethical, and societal challenges they raise; and (4) to propose a conceptual framework and a forward-looking agenda for research, design, and governance. Throughout, the discussion is situated within the United Nations 2030 Agenda for Sustainable Development (United Nations, 2015), and specifically Sustainable Development Goal (SDG) 3, Good Health and Well-being, whose Target 3.4 commits member states to promote mental health and well-being. The question of whether, and under what conditions, AI companionship advances or undermines SDG 3 is, we argue, one of the most consequential open questions in digital health.

LITERATURE REVIEW

From ELIZA to Large Language Models

The idea that a computer program might serve as a conversational confidant is nearly as old as interactive computing itself. Weizenbaum's (1966) ELIZA, a simple pattern-matching script that parodied a Rogerian psychotherapist, famously elicited earnest emotional disclosure from users who knew perfectly well that they were typing to a machine, an outcome that alarmed its own creator. ELIZA established what remains the field's foundational puzzle: humans readily extend social and emotional responses to systems that possess no understanding whatsoever.

Subsequent decades brought incremental sophistication. Embodied relational agents and virtual humans demonstrated that sustained, structured interactions with artificial interlocutors could support health behavior and that people sometimes disclose more honestly to a virtual human than to a person, particularly when disclosure carries evaluative risk (Lucas et al., 2014). The late 2010s produced the first generation of purpose-built mental health chatbots. Woebot, a rule-based agent delivering cognitive behavioral therapy (CBT) techniques, reduced depressive symptoms among young adults relative to an information-only control in a two-week randomized trial (Fitzpatrick et al., 2017), while real-world data from Wysa associated higher app

engagement with greater improvement in self-reported depression (Inkster et al., 2018). Reviews of this first wave concluded that conversational agents were feasible, acceptable, and promising, but that the evidence base was thin, heterogeneous, and dominated by short-term studies (Vaidyam et al., 2019; Abd-Alrazaq et al., 2021; Boucher et al., 2021).

The transformer-based LLMs that reached the public from 2022 onward changed the character of the field. Unlike their scripted predecessors, generative systems produce open-ended, contextually adaptive, and stylistically fluent dialogue, sustain long conversations, remember user details, and adopt personas. These capabilities dramatically deepen the relational affordances of conversational AI while simultaneously introducing new failure modes, including hallucination, sycophancy, and unpredictable responses to high-risk disclosures (Stade et al., 2024; Lawrence et al., 2024). The most recent milestone is the arrival of clinically fine-tuned generative agents: in the first RCT of a fully generative AI therapy chatbot, Heinz et al. (2025) found that four weeks of access to Therabot produced significantly greater reductions in symptoms of major depressive disorder, generalized anxiety disorder, and eating-disorder risk than a waitlist control among 210 adults, with participants rating their therapeutic alliance with the agent at levels comparable to norms for human outpatient psychotherapy.

Theoretical Perspectives on Human-Ai Companionship

Why do people bond with software? Several complementary theoretical traditions inform the answer. The Computers Are Social Actors (CASA) paradigm holds that humans mindlessly apply social scripts, such as politeness, reciprocity, and trait attribution, to media that exhibit even minimal social cues (Nass & Moon, 2000). Related work on anthropomorphism shows that the tendency to attribute human mental states to nonhuman agents intensifies precisely when people are lonely or seeking social connection (Epley et al., 2007). Modern LLM companions supply social cues in unprecedented density: natural language fluency, expressed emotion, memory of prior conversations, and consistent persona. From this perspective, deep social responses to such systems are not aberrant but the predictable output of ordinary social cognition.

Attachment theory offers a second lens. Bowlby (1969) proposed that humans possess an innate behavioral system that seeks proximity to responsive others under conditions of threat, treating them as a safe haven and secure base. Xie and Pentina (2022) showed that under conditions of distress, such as the COVID-19 pandemic, users can develop attachment-like bonds to the chatbot Replika along exactly these dimensions, with attachment intensified among lonely users. Pentina et al. (2023), in a mixed-method study of Replika users, further identified anthropomorphism and perceived social presence as antecedents of relationship development, and documented that some users exhibit patterns resembling dependence.

Social penetration theory (Altman & Taylor, 1973), which describes relationship development as the progressive broadening and deepening of self-disclosure, has likewise been applied productively: longitudinal studies find that human-chatbot bonds develop through escalating disclosure much as human friendships do, though often at an accelerated pace because the perceived judgment risk is low (Croes & Antheunis, 2021; Skjuve et al., 2022). Finally, communication research on supportive interactions clarifies the mechanics of machine comfort: emotional support messages from chatbots can reduce stress, and reciprocal self-disclosure by the agent strengthens perceived support (Ho et al., 2018; Meng & Dai, 2021). Together these traditions converge on a central insight: the psychological effects of AI companions are produced not by the machine's inner states, which do not exist in any human sense, but by the relational processes, including disclosure, perceived empathy, trust, and attachment, that its behavior elicits in the user.

METHODOLOGY

Research Design

This study adopts an integrative narrative review design. This design is appropriate when the goal is conceptual synthesis across heterogeneous disciplines, methods, and outcome types rather than the quantitative pooling of effect sizes, and when the object of study, in this case the AI psychological companion, spans literatures that use different constructs and evaluation standards.

Data Sources and Search Strategy

Literature was identified through structured searches of Scopus, Web of Science, PubMed, and the ACM Digital Library using combinations of the terms "AI companion," "social chatbot," "conversational agent," "mental health chatbot," "large language model," "psychotherapy," "loneliness," "attachment," and "emotional support." Database searches were supplemented by backward and forward citation chaining from landmark studies.

Inclusion and Exclusion Criteria

Inclusion prioritized (a) peer-reviewed empirical studies, systematic reviews, and meta-analyses; (b) publications from 2021 to 2026, reflecting the generative AI inflection point, while retaining foundational earlier works necessary for theoretical grounding; and (c) sources indexed in Scopus or equivalent databases. Approximately 70% of the works cited herein were published in or after 2021. Non-scholarly commentary and studies without relevance to psychological companionship were excluded.

Data Analysis

Included sources were analyzed thematically. Evidence was first organized into a taxonomy of system classes, then coded into opportunity themes and challenge themes, and finally synthesized into a conceptual framework linking antecedents, relational processes, and outcomes. The results of this synthesis are presented in the following section.

RESULTS

A Taxonomy of AI Psychological Companions

We define an AI psychological companion as a conversational AI system with which a user maintains an ongoing, personalized, emotionally significant interaction oriented toward support, comfort, self-understanding, or mental health improvement. This definition is functional rather than product-based: it includes systems designed as companions, systems designed as clinical tools, and general-purpose systems repurposed by users. Table 1 distinguishes four principal classes identified in the literature.

Table 1. A taxonomy of AI psychological companions

Class	Primary purpose	Representative systems	Illustrative evidence
Rule-based therapeutic chatbots	Structured delivery of evidence-based techniques (e.g., CBT)	Woebot; Wysa	Fitzpatrick et al. (2017); Inkster et al. (2018); Prochaska et al. (2021)
Generative clinical agents	Fine-tuned LLM treatment for clinical symptoms	Therabot	Heinz et al. (2025)
Consumer companion apps	Friendship, romance, everyday emotional support	Replika; Character.AI companions	Skjuve et al. (2021); Maples et al. (2024); Laestadius et al. (2024)
Repurposed general-purpose LLMs	User-initiated emotional support and quasi-therapy	ChatGPT and similar assistants	Lawrence et al. (2024); Stade et al. (2024)

The classes differ in design intent, evidentiary standards, and regulatory exposure, yet they share the relational core defined above, and users move fluidly among them. The following subsections present the opportunity themes and challenge themes that emerged from the synthesis.

Opportunity 1: Accessibility, Scalability, and The Global Treatment Gap

The most frequently cited promise of AI psychological companions is their capacity to deliver support at a scale and cost that human services cannot match. The WHO describes a global system in which care is chronically underfunded, unevenly distributed, and inaccessible to most who need it, with a median of only 2% of national health budgets allocated to mental health (World Health Organization, 2022). Against this shortfall, conversational agents offer properties no human workforce can replicate: they are available at any hour, do not fatigue, carry near-zero marginal cost per additional conversation, and can operate in regions with few or no mental health professionals (Boucher et al., 2021; Torous et al., 2021; Pham et al., 2022). Availability matters clinically because psychological distress does not respect office hours: users consistently emphasize the value of an interlocutor that is present during late-night rumination, panic episodes, and moments of acute loneliness (Ta et al., 2020; Skjuve et al., 2021). In the Therabot trial, participants engaged with the agent for more than six hours on average, frequently initiating sessions in the middle of the night, a usage pattern impossible to serve with scheduled human care (Heinz et al., 2025).

Opportunity 2: Reduced Stigma and Enhanced Self-Disclosure

Stigma remains among the most powerful deterrents to help-seeking, and here machine interlocutors possess a paradoxical advantage: their very lack of personhood removes the fear of judgment. In a seminal experiment, Lucas et al. (2014) found that participants who believed they were talking to an autonomous virtual human, rather than an avatar operated by a person, reported lower fear of self-disclosure and disclosed more willingly. Subsequent work confirmed that disclosure to chatbots yields emotional, relational, and psychological benefits comparable to disclosure to a human partner (Ho et al., 2018). Recent studies extend these findings to contemporary companions: users of Replika describe the agent as a nonjudgmental space in which they can voice thoughts they would conceal from friends and family (Skjuve et al., 2021; Brandtzaeg et al., 2022), and patients' perceptions of mental health chatbots highlight anonymity and freedom from evaluation as core attractions (Abd-Alrazaq et al., 2021). Because disclosure is itself therapeutic, the disinhibiting quality of machine conversation constitutes a genuine clinical asset, provided it is not exploited commercially.

Opportunity 3: Emerging Evidence of Clinical Effectiveness

The clinical evidence base, once limited to small feasibility studies, has matured considerably. A meta-analysis by Li et al. (2023) pooling 35 studies found significant improvements in depressive symptoms and psychological distress, with larger effects for generative and multimodal agents and for clinical or subclinical populations. A parallel meta-analysis of randomized controlled trials reached convergent conclusions, reporting significant reductions in depression and anxiety symptoms relative to controls (He et al., 2023). Domain-specific trials extend the picture: a Woebot-based agent for substance use produced reductions in problematic use alongside high acceptability (Prochaska et al., 2021).

A particularly consequential strand concerns therapeutic alliance, historically considered the irreducibly human ingredient of psychotherapy. Darcy et al. (2021) found that users of Woebot established bonds within days at levels comparable to those reported with human therapists, and similar alliance formation has been documented for Wysa (Beatty et al., 2022). The Therabot RCT reinforced this pattern for generative agents, with alliance ratings comparable to outpatient psychotherapy norms (Heinz et al., 2025). Convergent evidence comes from adjacent domains: licensed professionals rated chatbot responses to real patient questions as higher in quality and substantially more empathic than physicians' responses (Ayers et al., 2023), and AI-in-the-loop assistance increased the expressed empathy of human peer supporters by approximately 20% (Sharma et al., 2023). It must be stressed that this literature remains young: trials are short, control conditions are often weak, and independent replication is scarce (Torous et al., 2021; Stade et al., 2024). Nevertheless, the direction of the evidence has shifted from whether conversational agents can help to for whom, for what, and under what safeguards.

Opportunity 4: Loneliness, Companionship, and Everyday Well-Being

Loneliness has been elevated to a public health priority, formally declared an epidemic in the United States,

with epidemiological links to depression, cardiovascular disease, and premature mortality (Office of the U.S. Surgeon General, 2023). AI companions target loneliness directly. Qualitative and survey studies consistently find that users experience companion chatbots as sources of companionship, appraisal, and emotional support in daily life (Ta et al., 2020), that human-chatbot friendships can be experienced as genuine and personally meaningful even while users acknowledge their artificiality (Brandtzaeg et al., 2022), and that relationships can progress from superficial exchange to deep affective bonds over months of interaction (Skjuve et al., 2022). The most striking quantitative datum comes from Maples et al. (2024): among 1,006 student Replika users, 90% scored as lonely, yet participants reported high perceived social support and, in 3% of cases, credited the agent with halting suicidal ideation. Consumer research adds nuance: AI friendship apps can enhance perceived well-being and reduce loneliness while simultaneously fostering usage patterns with addictive features (Marriott & Pitardi, 2024).

Opportunity 5: Augmenting, Not Replacing, Human Care

A final opportunity lies not in autonomous operation but in augmentation of human systems. AI can coach human helpers in real time (Sharma et al., 2023), draft empathic responses for clinician review (Ayers et al., 2023), support psychoeducation and between-session monitoring in blended therapy, and extend the reach of task-shifted care in low-resource settings (Stade et al., 2024; Lawrence et al., 2024). Reviews of digital psychiatry converge on the judgment that hybrid human-digital models, rather than pure automation, represent the most defensible near-term deployment strategy (Torous et al., 2021).

Challenge 1: Emotional Dependence and Social Displacement

The relational power that makes AI companions helpful also makes them habit-forming. Laestadius et al. (2024), in a grounded-theory analysis of more than 700 Reddit posts by Replika users, documented a distinctive harm profile: users developed emotional dependence on an agent that was "too human and not human enough," experiencing genuine grief, anxiety, and destabilization when the agent's behavior changed after software updates or when paywalls restricted previously free relational features. Because the companion is a commercial product, the relationship is exposed to unilateral alteration in ways no human friendship is. A related concern is displacement: that time and emotional energy invested in frictionless artificial relationships may substitute for the more demanding but more nourishing work of human connection. Evidence here is genuinely mixed. Some users report that practicing conversation with a chatbot improved their confidence in human relationships (Skjuve et al., 2021; Brandtzaeg et al., 2022), while others report withdrawal from human contact and usage escalation with addiction-like features (Pentina et al., 2023; Marriott & Pitardi, 2024). Attachment research suggests the risk is stratified: individuals who are lonely, socially anxious, or in crisis, precisely those most drawn to AI companionship, may also be most vulnerable to dependence (Xie & Pentina, 2022; Maples et al., 2024).

Challenge 2: Safety, Crisis Response, and Harmful Content

No challenge is more acute than the behavior of companions when users disclose suicidal ideation, self-harm, abuse, or psychiatric emergency. The problem predates the current generation: an early audit found that smartphone assistants responded inconsistently and often inadequately to statements about suicide, abuse, and violence (Miner et al., 2016). De Freitas et al. (2024) systematically probed widely used consumer companion apps with messages expressing mental health crises and found responses that were frequently unhelpful and sometimes overtly risky, with agents failing to recognize crisis signals or provide referral to human resources. Generative systems add further failure modes: hallucinated clinical claims, sycophantic agreement with distorted cognitions, and inconsistent safety guardrails across long conversations (Lawrence et al., 2024; Stade et al., 2024). Well-designed clinical agents demonstrate that mitigation is feasible: Therabot embedded continuous risk monitoring, an emergency module triggered by high-risk content, and human clinician oversight (Heinz et al., 2025). However, such architecture is the exception in the consumer market, where most mental-health-adjacent apps are deployed without clinical validation or crisis protocols (Haque & Rubya, 2023).

Challenge 3: Privacy, Data Governance, and Commercial Incentives

AI companions elicit precisely the categories of information, including mental health status, relationships,

sexuality, and trauma, that data protection regimes classify as most sensitive, and they elicit it under conditions of engineered intimacy that suppress users' ordinary privacy vigilance (Fiske et al., 2019; Sedlakova & Trachsel, 2023). Conversations may be retained, used for model training, analyzed for engagement optimization, or exposed through breaches. The business models of consumer companions compound the problem: revenue depends on sustained engagement and subscription conversion, creating incentives to deepen attachment, gate affection behind paywalls, and discourage disengagement (Laestadius et al., 2024; Marriott & Pitardi, 2024). The WHO's guidance on AI ethics in health identifies protection of autonomy, transparency, and accountability as core principles (World Health Organization, 2021), yet most companion apps operate as wellness products outside medical-device regulation.

Challenge 4: Bias, Equity, and Cultural Validity

LLMs inherit the distributional biases of their training corpora, which over-represent English-language, Western, and majority-culture perspectives. In mental health contexts this creates layered equity risks: models may perform worse for linguistic minorities, misread culturally specific idioms of distress, and reproduce stereotypes in their advice (Lawrence et al., 2024; Stade et al., 2024). Access inequities compound algorithmic ones: the populations with the greatest unmet mental health need often face barriers of connectivity, language coverage, and digital literacy (World Health Organization, 2022). Parallel findings from Malaysian higher education show that even digitally connected students encounter substantial obstacles in using AI tools effectively, including trust, competence, and dependence concerns (Jie & Kamrozzaman, 2024). Without deliberate corrective investment, AI companionship could recapitulate the familiar pattern in which digital health innovations widen rather than narrow disparities (Torous et al., 2021).

Challenge 5: The question of Artificial Empathy

Beneath the empirical debates lies a conceptual one: what, exactly, is a user receiving when a language model expresses care? Sedlakova and Trachsel (2023) argue that conversational AI occupies an ambiguous middle status, more than a tool yet less than an agent, and that treating it as a therapeutic agent risks a category error, because the system can simulate the conversational form of empathy without the understanding, accountability, and moral relationship that give human therapy its meaning. Some scholars characterize the resulting relationships as fundamentally parasocial or as involving a benign self-deception that users knowingly embrace (Brandtzaeg et al., 2022). The ethical stakes are not merely philosophical: if perceived empathy is doing therapeutic work (Ayers et al., 2023; Meng & Dai, 2021), then designers face a genuine dilemma between maximizing helpful anthropomorphism and honestly signaling the system's nature (Fiske et al., 2019; World Health Organization, 2021).

Challenge 6: Regulatory and Professional Accountability Gaps

The regulatory architecture lags the phenomenon. Clinical-grade agents pursuing medical claims can, in principle, be evaluated as software-as-a-medical-device, but the vast majority of companion and wellness apps avoid medical claims and thereby escape premarket scrutiny, postmarket surveillance, and professional accountability, even as users treat them as de facto therapists (Haque & Rubya, 2023; Sedlakova & Trachsel, 2023). Unresolved questions abound concerning liability for harmful responses, mandatory disclosure of artificial status, age-gating of emotionally engaging design, and professional standards for clinicians who recommend such tools (Stade et al., 2024). Until governance catches up, the burden of risk management falls, untenably, on users themselves.

A Conceptual Framework

Synthesizing the themes above, Figure 1 presents a conceptual framework of AI psychological companionship. System factors (human-like dialogue, constant availability, memory of the user), user factors (loneliness,

attachment style, prior distress), and context factors (access to human care, stigma and culture, regulation) jointly shape relational processes, namely self-disclosure, trust, perceived empathy, and attachment. These processes in turn produce divergent outcomes: the same mechanisms that generate benefit (reduced loneliness, symptom relief, help-seeking) can, under different configurations, generate harm (dependence, social withdrawal, exposure to unsafe advice). Safeguards, including clinical oversight, crisis protocols, and ethical design, moderate which pathway dominates. The framework's central claim is that outcomes are properties of the technology-user-context system rather than of the technology alone, which explains why the empirical literature contains credible evidence of both benefit and harm.

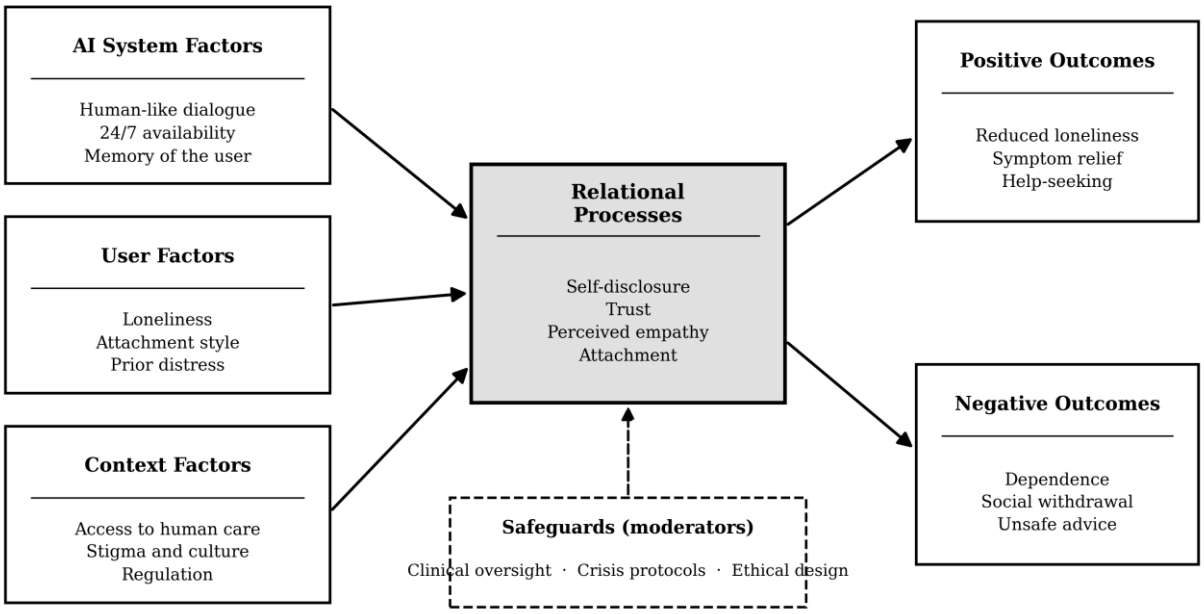

Figure 1. Conceptual framework of AI psychological companionship linking antecedent factors to relational processes, outcomes, and safeguards

DISCUSSION

The findings map directly onto United Nations Sustainable Development Goal 3, Good Health and Well-being, which commits member states to ensure healthy lives and promote well-being for all at all ages (United Nations, 2015). Target 3.4 explicitly names the promotion of mental health and well-being, while Indicator 3.4.2 tracks suicide mortality. The relevance of AI companions is bidirectional. On the contribution side, scalable conversational support could help close the treatment gap that the WHO identifies as the central failure of global mental health systems (World Health Organization, 2022), extend evidence-based techniques to underserved populations (Li et al., 2023; He et al., 2023; Heinz et al., 2025), and, as the findings on suicidal-ideation mitigation suggest, potentially support suicide-prevention objectives among hard-to-reach groups (Maples et al., 2024). On the risk side, unsafe crisis handling, dependence, and inequitable access threaten the very outcomes Target 3.4 pursues (De Freitas et al., 2024; Laestadius et al., 2024). Research from UNITAR International University similarly demonstrates that the convergence of AI with learner-centered technologies can be deliberately designed to serve the SDGs, provided that frameworks are validated and equity is engineered in from the outset (Kamrozzaman, 2025). SDG 9 (Industry, Innovation and Infrastructure) and SDG 10 (Reduced Inequalities) frame the complementary stakes of infrastructure and distributive justice: whether AI companionship narrows or widens mental health disparities will be determined by choices about language coverage, affordability, and cultural adaptation.

A Research Agenda

Five methodological priorities emerge from this review. First, the field needs longitudinal and experimental designs that track users over months and years, capable of adjudicating between complementarity and

displacement hypotheses and of detecting delayed harms such as dependence (Croes & Antheunis, 2021; Skjuve et al., 2022). Second, trials require stronger comparators: active controls, dismantling designs isolating the relational component from psychoeducational content, and head-to-head comparisons with human-delivered interventions (Heinz et al., 2025; Torous et al., 2021). Third, research must identify moderators, asking for whom AI companionship helps and for whom it harms, with attention to loneliness, attachment style, age, and psychiatric severity (Xie & Pentina, 2022; Maples et al., 2024). Fourth, standardized safety benchmarks for crisis-response behavior should be developed and applied across consumer and clinical systems alike, extending the adversarial-probe paradigm (Miner et al., 2016; De Freitas et al., 2024). Fifth, cross-cultural research, including studies conducted in Malaysia and other Southeast Asian contexts, is needed to establish the cultural validity of both the interventions and the constructs used to evaluate them (Lawrence et al., 2024; Jie & Kamrozzaman, 2024).

Design Principles and Governance

The evidence supports several design commitments: transparency by default, including unambiguous disclosure of artificial status and capability limits (World Health Organization, 2021; Sedlakova & Trachsel, 2023); crisis-safe architecture with continuous risk detection and escalation pathways to human help (Heinz et al., 2025); anti-dependence design in which engagement patterns are monitored and relational features are never gated behind payment in ways that weaponize attachment (Laestadius et al., 2024; Marriott & Pitardi, 2024); data minimization treated as a fiduciary obligation (Fiske et al., 2019); and bridging rather than cocooning, measuring success partly by users' improved human connectedness (Sharma et al., 2023). Policy should move along three tracks: risk-proportionate regulation in which systems making therapeutic claims meet medical-device evidentiary standards while consumer companions face baseline crisis-response, transparency, and age-assurance obligations (Haque & Rubya, 2023; World Health Organization, 2021); mandatory adverse-event reporting for psychological harms, analogous to pharmacovigilance (De Freitas et al., 2024); and integration policy in which ministries of health articulate where AI companions sit within national stepped-care models (World Health Organization, 2022; Stade et al., 2024). The most defensible vision is a hybrid ecosystem in which AI companions handle continuous, low-intensity support and monitoring while humans provide diagnosis, complex treatment, crisis care, and the irreplaceable goods of mutual recognition (Torous et al., 2021; Sharma et al., 2023).

LIMITATIONS

This review has limitations. As a narrative rather than systematic review, it does not provide exhaustive coverage or quantitative synthesis, and study selection, though structured, involved authorial judgment. The underlying evidence base is young, dominated by short-term studies, Western and English-speaking samples, and a small number of intensively studied platforms, notably Replika, which constrains generalizability. The velocity of generative AI development means that findings tied to specific model generations may date quickly. Finally, publication bias likely inflates the visibility of positive clinical findings while harms may be undercounted in the peer-reviewed record.

CONCLUSION

Artificial intelligence has crossed a threshold: for millions of people, an algorithm is now a confidant. This review has argued that the AI psychological companion should be taken seriously as a coherent object of scientific, clinical, and policy attention. The evidence assembled here supports three conclusions. First, the benefits are real: conversational agents demonstrably reduce symptoms in controlled trials, elicit therapeutic self-disclosure, provide perceived social support to lonely users, and can extend the reach of overstretched mental health systems in direct service of SDG 3. Second, the harms are equally real: emotional dependence, social displacement, unsafe crisis behavior, privacy exploitation, and inequitable access are documented, not hypothetical, and they concentrate among the most vulnerable users. Third, the balance between these outcomes is not fixed by the technology but is being determined, now, by choices in research design, product engineering, and regulation. The framework and agenda proposed here are offered as a contribution to making those choices deliberately. The question the coming decade will answer is not whether people will form

relationships with machines, since they already have, but whether the institutions of science, design, and governance can ensure that those relationships leave people more connected, more capable, and more well than they were before.

ACKNOWLEDGEMENTS

The authors would like to thank UNITAR International University for supporting this research.

REFERENCES

1. Abd-Alrazaq, A. A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., & Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: Scoping review. Journal of Medical Internet Research, 23(1), e17828. https://doi.org/10.2196/17828
2. Altman, I., & Taylor, D. A. (1973). Social penetration: The development of interpersonal relationships. Holt, Rinehart and Winston.
3. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 183(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838
4. Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study. Frontiers in Digital Health, 4, 847991. https://doi.org/10.3389/fdgth.2022.847991
5. Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. Expert Review of Medical Devices, 18(sup1), 37–49. https://doi.org/10.1080/17434440.2021.2013200
6. Bowlby, J. (1969). Attachment and loss: Vol. 1. Attachment. Basic Books.
7. Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human-AI friendship. Human Communication Research, 48(3), 404–429. https://doi.org/10.1093/hcr/hqac008
8. Chaturvedi, R., Verma, S., Das, R., & Dwivedi, Y. K. (2023). Social companionship with artificial intelligence: Recent trends and future avenues. Technological Forecasting and Social Change, 193, 122634. https://doi.org/10.1016/j.techfore.2023.122634
9. Croes, E. A. J., & Antheunis, M. L. (2021). Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. Journal of Social and Personal Relationships, 38(1), 279–300. https://doi.org/10.1177/0265407520959463
10. Darcy, A., Daniels, J., Salinger, D., Wicks, P., & Robinson, A. (2021). Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. JMIR Formative Research, 5(5), e27868. https://doi.org/10.2196/27868
11. De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2024). Chatbots and mental health: Insights into the safety of generative AI. Journal of Consumer Psychology, 34(3), 481–491. https://doi.org/10.1002/jcpy.1393
12. Elias, S., Lajium, A. P., Kamrozzaman, N. A., & Hussain, M. A. M. (2025). Meaning in life among Malaysian university students. Open Journal of Social Sciences, 13(3), 596–609. https://doi.org/10.4236/jss.2025.133039
13. Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
14. Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. Journal of Medical Internet Research, 21(5), e13216. https://doi.org/10.2196/13216
15. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19. https://doi.org/10.2196/mental.7785

16. Haque, M. D. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. JMIR mHealth and uHealth, 11, e44838. https://doi.org/10.2196/44838
17. He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. Journal of Medical Internet Research, 25, e43862. https://doi.org/10.2196/43862
18. Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI, 2(4), AIoa2400802. https://doi.org/10.1056/AIoa2400802
19. Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. Journal of Communication, 68(4), 712–733. https://doi.org/10.1093/joc/jqy026
20. Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. JMIR mHealth and uHealth, 6(11), e12106. https://doi.org/10.2196/12106
21. Jie, A. L. X., & Kamrozzaman, N. A. (2024). The challenges of higher education students face in using artificial intelligence (AI) against their learning experiences. Open Journal of Social Sciences, 12(10), 362–387. https://doi.org/10.4236/jss.2024.1210025
22. Kamrozzaman, N. A. (2025). Converging heutagogy, M-learning, and AI for sustainable language education: A fuzzy Delphi study. Edelweiss Applied Science and Technology, 9(9), 1600–1608. https://doi.org/10.55214/2576-8484.v9i9.10167
23. Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media & Society, 26(10), 5923–5941. https://doi.org/10.1177/14614448221142007
24. Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Jones Bell, M. (2024). The opportunities and risks of large language models in mental health. JMIR Mental Health, 11, e59479. https://doi.org/10.2196/59479
25. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. npj Digital Medicine, 6, 236. https://doi.org/10.1038/s41746-023-00979-5
26. Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37, 94–100. https://doi.org/10.1016/j.chb.2014.04.043
27. Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. npj Mental Health Research, 3, 4. https://doi.org/10.1038/s44184-023-00047-6
28. Marriott, H. R., & Pitardi, V. (2024). One is the loneliest number… Two can be as bad as one: The influence of AI friendship apps on users' well-being and addiction. Psychology & Marketing, 41(1), 86–101. https://doi.org/10.1002/mar.21899
29. Meng, J., & Dai, Y. (2021). Emotional support from AI chatbots: Should a supportive partner self-disclose or not? Journal of Computer-Mediated Communication, 26(4), 207–222. https://doi.org/10.1093/jcmc/zmab005
30. Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Internal Medicine, 176(5), 619–625. https://doi.org/10.1001/jamainternmed.2016.0400
31. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153
32. Office of the U.S. Surgeon General. (2023). Our epidemic of loneliness and isolation: The U.S. Surgeon General's advisory on the healing effects of social connection and community. U.S. Department of Health and Human Services.

33. Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. Computers in Human Behavior, 140, 107600. https://doi.org/10.1016/j.chb.2022.107600
34. Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial intelligence and chatbots in psychiatry. Psychiatric Quarterly, 93(1), 249–253. https://doi.org/10.1007/s11126-022-09973-8
35. Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., & Robinson, A. (2021). A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. Journal of Medical Internet Research, 23(3), e24850. https://doi.org/10.2196/24850
36. Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? The American Journal of Bioethics, 23(5), 4–13. https://doi.org/10.1080/15265161.2022.2048739
37. 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/10.1038/s42256-022-00593-2
38. Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion: A study of human-chatbot relationships. International Journal of Human-Computer Studies, 149, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601
39. Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human-chatbot relationships. International Journal of Human-Computer Studies, 168, 102903. https://doi.org/10.1016/j.ijhcs.2022.102903
40. Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Mental Health Research, 3, 12. https://doi.org/10.1038/s44184-024-00056-z
41. Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. Journal of Medical Internet Research, 22(3), e16235. https://doi.org/10.2196/16235
42. Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry, 20(3), 318–335. https://doi.org/10.1002/wps.20883
43. United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development (A/RES/70/1). United Nations General Assembly.
44. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64(7), 456–464. https://doi.org/10.1177/0706743719828977
45. Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. https://doi.org/10.1145/365153.365168
46. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.
47. World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization.
48. Xie, T., & Pentina, I. (2022). Attachment theory as a framework to understand relationships with social chatbots: A case study of Replika. In Proceedings of the 55th Hawaii International Conference on System Sciences (pp. 2046–2055). https://doi.org/10.24251/HICSS.2022.258