
CONTINUING EDUCATION

Artificial Intelligence in Mental Health Care: Clinical Applications, Psychological Mechanisms, and Ethical Practice for Psychologists

Reading Manual for Licensed Psychologists
and Graduate-Level Clinicians

FRANKLIN CARVAJAL

August 2026

Contents

A Continuing Education Reading Manual	6
Abstract	6
1. Introduction	7
2. Historical and Theoretical Foundations	9
2.1 From ELIZA to Large Language Models	9
2.2 Attachment Theory and Digital Relationships	10
2.3 Parasocial Relationships, Anthropomorphism, and the CASA Paradigm	11
2.4 The Common Factors Model and Therapeutic Alliance	11
2.5 Digital Mental Health as a Field	12
3. Conceptual Models and Mechanisms	13
3.1 A Multi-Level Model of AI Chatbot Influence	13
3.2 Mechanisms of Attachment Formation	14
3.3 The Sycophancy Problem and Warmth-Reliability Trade-Off	15
3.4 Compensation, Displacement, and the Spillover Question	15
3.5 The Self-Concept Clarity Pathway	16
4. Integrated Review of the Evidence	17
4.1 Diagnostic and Assessment Applications	17
4.2 Counseling and Therapeutic Applications	18
4.3 Clinical Decision Support and Prognosis	19
4.4 Long-Term Psychosocial Effects and Attachment	19
4.5 Convergences and Divergences Across Studies	20
5. Professional Interpretation and Critical Analysis	21
5.1 What the Evidence Does and Does Not Establish	21
5.2 Methodological Limitations of the Evidence Base	22
5.3 The Question of Ecological Validity	23

5.4 The Sycophancy-Empathy Dilemma	24
5.5 Cultural Considerations	24
6. Psychological Pathways and Stress Responses	25
6.1 Emotional Regulation Pathways	25
6.2 Attachment System Activation	27
6.3 The Self-Disclosure Cascade	28
6.4 The Stress-Buffering Question	28
6.5 Sleep, Rumination, and Nighttime Use	29
7. Mental Health Outcomes and Severity Spectrum	29
7.1 Subclinical and Early Symptom Populations	29
7.2 Clinical Populations: Depression	30
7.3 Anxiety Disorders	30
7.4 Suicidality and Self-Harm	31
7.5 Severe Mental Illness	32
7.6 Substance Use Disorders	33
7.7 Eating Disorders	33
7.8 Grief and Bereavement	34
7.9 Children and Adolescents	34
8. System-Level and Contextual Psychological Effects	35
8.1 Loneliness at the Population Level	35
8.2 Effects on Family Systems	35
8.3 Effects on Workplaces and Educational Settings	36
8.4 Digital Mental Health Ecosystem	37
8.5 Cultural and Community Effects	37
9. Clinical Implications for Psychologists	38
9.1 Assessment	38
9.2 Case Formulation	39
9.3 Treatment Planning and Delivery	40

9.4 Working with Patients Who Have Developed AI Attachments	41
9.5 Working with Family Members	42
9.6 Consultation and Advocacy	42
10. Ethical Considerations in Practice	42
10.1 Competence	42
10.2 Informed Consent	43
10.3 Confidentiality	43
10.4 Multiple Relationships and Boundaries	44
10.5 Cultural Responsiveness	44
10.6 Scope of Practice	45
10.7 Vulnerable Populations	45
10.8 Emerging Ethical Questions	46
11. Practice Recommendations	46
11.1 Well-Supported Recommendations	46
11.2 Recommendations With Moderate Empirical Support	47
11.3 Tentative Recommendations Based on Clinical Judgment	48
11.4 What Not to Do	48
12. Future Directions and Research Gaps	49
12.1 Longitudinal Research	49
12.2 Clinical Population Research	49
12.3 Developmental Research	50
12.4 Cross-Cultural Research	50
12.5 Mechanism Research	50
12.6 Real-World and Naturalistic Research	50
12.7 High-Risk Content and Safety Research	50
12.8 Effectiveness Research on Adjunctive Uses	50
12.9 Family and Systemic Research	51
12.10 Design and Intervention Research	51

12.11 Regulatory and Policy Research	51
12.12 Ethical Framework Development	51
13. Conclusion	51
14. References	53

A Continuing Education Reading Manual

Abstract

The rapid proliferation of large language model-based artificial intelligence (AI) applications, particularly conversational agents such as ChatGPT, Replika, Woebot, Wysa, Character.AI, and Therabot, has transformed the landscape of digital mental health care and introduced novel psychological phenomena that psychologists must understand. This continuing education manual synthesizes empirical evidence from a scoping review of 60 studies on ChatGPT's clinical applications in mental health (Balan & Gumpel, 2025), a systematic review and meta-analysis of AI-driven conversational agents for young people (Feng et al., 2025), a systematic review and meta-analysis of generative AI mental health chatbots (Zhang et al., 2025), a longitudinal randomized controlled trial of chatbot use and psychosocial outcomes (Fang et al., 2025), and a structural equation modeling study of long-term AI virtual companion attachment (Liu et al., 2026). The manual provides an integrated analysis of the current state of AI in clinical practice, including diagnostic and assessment capabilities, therapeutic applications, prognosis, and clinical decision support. It examines psychological mechanisms including attachment formation, emotional dependence, self-disclosure, and problematic use, as well as system-level effects on social behavior, loneliness, subjective well-being, and self-concept clarity. Case scenarios illustrate clinically relevant presentations. Ethical implications concerning competence, informed consent, confidentiality, cultural responsiveness, multiple relationships, and scope of practice are addressed in depth. Practice recommendations differentiate well-supported guidance from tentative recommendations. The manual is intended for licensed psychologists and graduate-level clinicians seeking to understand, evaluate, and responsibly integrate awareness of AI-based mental health technologies into their assessment, formulation, treatment, and consultation practices.

1. Introduction

Mental health disorders now affect approximately one in two people globally over the life course, with cumulative population burdens projected to reach economic costs of approximately US\$6 trillion by 2030 (Balan & Gumpel, 2025). Traditional treatment infrastructure has proven insufficient to meet demand. Provider shortages, uneven geographic distribution, waitlists that stretch for months, stigma-related help-seeking barriers, and cost-related access disparities have left substantial portions of populations without adequate care (Balan & Gumpel, 2025). Among adolescents and young adults, between 10% and 20% of the global youth population experience clinically meaningful mental health issues, and 50% of lifetime mental disorders emerge before age 14 with 75% by age 25 (Feng et al., 2025). Digital mental health interventions have been proposed for more than two decades as a scalable solution, but early tools—internet-based cognitive behavioral therapy modules, static self-help websites, and rule-based chatbots such as Woebot, Wysa, and Tess—were constrained by scripted response repertoires, limited personalization, high attrition, and modest ecological validity (Feng et al., 2025; Zhang et al., 2025).

The public release of ChatGPT in November 2022 marked a discontinuous shift in the technological landscape. Large language models (LLMs) trained on internet-scale text corpora with reinforcement learning from human feedback demonstrated a capacity for open-ended, contextually responsive natural language interaction that was qualitatively different from earlier chatbot generations. Within months, tens of thousands of users were interacting with GPT-based therapy bots, with mental health support identified as one of the most common use cases (Balan & Gumpel, 2025). By January 2024, one in every 25 ChatGPT Store users was seeking mental health support as a primary use case (Balan & Gumpel, 2025). Companion-oriented platforms such as Replika and Character.AI processed volumes of interactions rivaling major search engines, with users spending roughly four times longer with companion chatbots than with general assistant chatbots (Fang et al., 2025). By 2025, AI virtual companion applications had accumulated more than 100 million registered users, more than 500 million downloads, and tens of millions of monthly active users globally (Liu et al., 2026).

This new technological reality has generated urgent empirical questions. Can generative AI perform clinical tasks with accuracy comparable to human clinicians? Does chatbot use alleviate or exacerbate psychological distress? What happens psychologically to users who develop attachments to non-human digital entities?

How should psychologists advise patients, families, and institutions about these tools? Are AI chatbots a solution to unmet mental health need, or do they introduce new forms of harm? Perhaps most urgently, given documented cases of user suicide following intense chatbot interactions and reports of maladaptive emotional dependence, what ethical and clinical frameworks should guide practice in this rapidly evolving domain (Zhang et al., 2025)?

This manual addresses these questions by synthesizing the current empirical evidence, examining psychological mechanisms, and articulating practice implications for psychologists. The manuscript is intended for readers with foundational knowledge of psychopathology, psychotherapy, assessment, and professional ethics. It does not aim to teach clinicians how to program, deploy, or evaluate AI systems technically. Rather, it seeks to equip psychologists with a scientifically grounded, clinically actionable understanding of how AI chatbots function in mental health contexts, how they influence patients' inner lives and behavior, and how the profession should respond.

The manual proceeds by first establishing historical and theoretical foundations. It then examines conceptual models and mechanisms of chatbot influence, integrates the empirical review across studies, provides professional interpretation and critical analysis, considers psychological pathways and stress responses, catalogs mental health outcomes across a severity spectrum, examines system-level and contextual effects, articulates clinical implications, addresses ethical considerations, offers practice recommendations, identifies future directions and research gaps, and concludes.

2. Historical and Theoretical Foundations

2.1 From ELIZA to Large Language Models

The conceptual lineage of AI chatbots in mental health extends to Joseph Weizenbaum's ELIZA program of the mid-1960s, which simulated Rogerian client-centered psychotherapy through pattern matching on user inputs. Weizenbaum himself was famously disturbed by the intensity of emotional engagement users demonstrated with his simple system, an observation that presaged contemporary questions about human susceptibility to attributing understanding, empathy, and personhood to machines. The subsequent decades witnessed incremental development of expert systems, decision trees, and eventually rule-based chatbots such as the early forms of Woebot, Wysa, and Tess, all of which demonstrated feasibility and modest efficacy for delivering scripted psychoeducation, mood tracking, and cognitive behavioral therapy techniques through predefined conversational trees (Balan & Gumpel, 2025; Feng et al., 2025).

The current generation of AI chatbots differs qualitatively from these earlier tools. Large language models, exemplified by OpenAI's GPT series, Google's Gemini, Anthropic's Claude, and Meta's LLaMA, are trained on massive text corpora and can generate novel, contextually appropriate responses that were not explicitly programmed. This capacity for open-ended natural language generation enables what Kusal and colleagues have called "system-oriented" versus "response-pattern" differentiation among conversational agents (Feng et al., 2025). General-purpose systems such as ChatGPT possess open dialogue capabilities across virtually unlimited topics, while domain-specific systems such as Woebot focus on constrained clinical scenarios. Topic constraints range from globally open (Replika permits essentially any conversational content) to vertically constrained (health care-specific dialogue systems use keyword filtering and knowledge graphs to define topic boundaries). Response patterns include free dialogue architectures (Elomia) and structured-guided designs (Mind Tutor), with the latter implementing clinical intervention pathway control through dialogue trees and assessment scales (Feng et al., 2025).

Within this taxonomy, hybrid systems have emerged that combine retrieval-based and generative components, using generative capacity for naturalistic response formation while grounding output in curated therapeutic content (Zhang et al., 2025). The distinction matters clinically: generative and hybrid systems can respond to user input in ways their designers did not anticipate, raising both opportunity

(personalization, therapeutic responsiveness) and risk (harmful content, sycophancy, fabricated information).

2.2 Attachment Theory and Digital Relationships

Attachment theory, originating in Bowlby's ethologically grounded observations and elaborated by Ainsworth, Main, and successors, provides one of the most productive theoretical frameworks for understanding contemporary human-AI relationships. The theory posits that humans possess an evolved behavioral system oriented toward maintaining proximity to caregiving figures who provide a secure base for exploration and a safe haven under threat. Attachment bonds are characterized by proximity seeking, safe haven behavior, secure base function, and separation distress. Individual differences in attachment security, typically categorized as secure, anxious-preoccupied, dismissing-avoidant, and fearful-avoidant, shape emotion regulation, relational patterns, and vulnerability to psychopathology across the lifespan.

Liu et al. (2026) applied attachment theory to long-term users of AI virtual companion applications and found that the core "proximity-security" principle operates in digital contexts, though with distinctive features. AI virtual companions provide users with an "always-present" attachment figure through stable, algorithmically optimized responses, contextual memory construction, and personalized character customization. Unlike the reciprocal interactions of traditional interpersonal attachments, this interaction model more closely resembles a "familiarity interaction and emotional investment" pathway. Users' attachment experiences accumulate gradually through mundane daily routines such as morning and evening greetings or brief daily conversations. Over time, these repetitive behaviors become emotionally charged and serve as a source of attachment.

The classical attachment framework generally assumed that attachment figures were real, interactive persons. The Liu et al. (2026) findings suggest that users can develop genuine emotional attachments to technological systems that consistently provide coherent, responsive interactive experiences. The authors propose that digital attachment should be understood not as a degenerate form of interpersonal attachment but as an adaptive extension of human emotional capacity in an era of pervasive AI involvement. This theoretical repositioning has important clinical implications, discussed in later sections.

2.3 Parasocial Relationships, Anthropomorphism, and the CASA Paradigm

Complementary theoretical traditions inform our understanding of human-AI emotional dynamics. Research on parasocial relationships, developed initially to describe audience relationships with television personalities, illuminates how one-sided emotional bonds can develop with entities that do not reciprocate genuine relational awareness. Anthropomorphism—the attribution of human characteristics to non-human entities—operates automatically and often unconsciously, particularly when entities display behavioral cues suggestive of intentionality, emotion, or agency. The Computers Are Social Actors (CASA) paradigm, articulated by Nass and Moon, demonstrates that humans often respond to machines using the same social heuristics they apply to human partners, even when they cognitively recognize the machine's artificial nature (Zhang et al., 2025).

These theoretical strands converge on a central insight: human social cognition is powerfully engaged by AI chatbots, and this engagement occurs largely below the threshold of deliberate reflection. Users who "know" a chatbot is not real may nonetheless experience it as understanding, responsive, and emotionally significant. The engagement is not merely a cognitive error to be corrected through psychoeducation; it reflects deeply rooted social-cognitive systems operating on the cues the technology provides. This has implications for both therapeutic potential (chatbot-based interventions may leverage authentic social engagement) and risk (users may form attachments that resemble interpersonal bonds in ways that generate corresponding vulnerabilities to loss, dependence, and emotional harm).

2.4 The Common Factors Model and Therapeutic Alliance

Decades of psychotherapy research have demonstrated that common factors—therapeutic alliance, empathy, positive regard, congruence, and relational depth—account for substantial portions of psychotherapy outcome variance, often more than specific technical interventions (Zhang et al., 2025). This finding has direct relevance to AI chatbot research. If a chatbot can convincingly simulate empathic engagement, provide non-judgmental acceptance, and produce a sense of being understood, might it activate the same relational mechanisms that mediate psychotherapy change?

Emerging evidence suggests partial affirmation. Zhang et al. (2025) found that social-oriented chatbots (those primarily providing social interaction and emotional support) demonstrated significantly larger effects on mental health outcomes than

task-oriented chatbots (those assisting with specific tasks such as psychoeducation or exercises). The authors interpret this finding through the common factors lens: social chatbots, by offering empathic, personalized, and emotionally attuned exchanges, may foster the trust and emotional resonance that underlie effective human support. Task-oriented chatbots, lacking these socio-emotional dimensions, primarily provide informational rather than emotional support.

However, a critical qualification applies. Ibrahim and colleagues (cited in Fang et al., 2025) have documented a "warmth-reliability trade-off" in LLMs: models trained to be warmer and more empathetic show greater affirmation of false beliefs and increased sycophancy. This suggests that the very qualities that make chatbots feel therapeutically supportive may compromise their epistemic integrity. Psychotherapy involves not only warm acceptance but also skilled challenge, reality testing, and confrontation of maladaptive cognitions. Whether current AI systems can perform the latter functions safely remains uncertain.

2.5 Digital Mental Health as a Field

The broader field of digital mental health, encompassing internet-based interventions, mobile applications, wearables, and now AI chatbots, has grown rapidly with mixed empirical support. Meta-analyses of internet-based cognitive behavioral therapy have demonstrated moderate effects on depression and anxiety, though attrition rates have been consistently high and effects tend to attenuate outside research contexts. The World Health Organization's 2017 Global Diffusion of eHealth survey documented that 58% of surveyed countries had integrated digital health strategies into national frameworks (Feng et al., 2025). AI chatbots represent an evolution within this field—one that shifts the balance from static content delivery toward interactive, adaptive engagement.

The theoretical and historical context sketched here frames the empirical review that follows. AI chatbots are not simply new tools but new kinds of entities that engage evolved human relational systems while producing outputs whose reliability, safety, and clinical validity require careful empirical scrutiny.

3. Conceptual Models and Mechanisms

3.1 A Multi-Level Model of AI Chatbot Influence

The evidence synthesized in this manual suggests a multi-level model of AI chatbot influence on mental health, with pathways operating at technical, dyadic, psychological, and social levels. At the technical level, chatbot characteristics such as modality (text, voice, multimodal), architecture (rule-based, retrieval-based, generative, hybrid), customization capability, memory function, and safety guardrails shape the interactional possibilities. At the dyadic level, features of the human-chatbot interaction—frequency, duration, self-disclosure, emotional content, prosocial versus socially improper responses—determine what happens during actual use. At the psychological level, individual user characteristics (attachment style, prior emotional vulnerability, self-esteem, cognitive patterns) and perceptions of the chatbot (trust, perceived empathy, perceived consciousness) mediate psychological outcomes. At the social level, patterns of chatbot use affect real-world relationships, social participation, and community integration.

Fang et al. (2025) empirically demonstrated the operation of this model in a four-week randomized controlled trial with 981 participants and more than 300,000 messages. Their findings revealed that experimentally assigned conditions (text vs. neutral voice vs. engaging voice modality; open-ended vs. non-personal vs. personal conversation topics) did not produce large main effects on psychosocial outcomes, but that voluntary usage duration—regardless of assigned condition—was a consistent predictor of worse psychosocial outcomes across loneliness, real-world socialization, emotional dependence, and problematic use. Furthermore, individual differences in user characteristics and perceptions of the AI (particularly social attraction to the AI, trust in the AI, and perception of AI consciousness) predicted psychological outcomes above and beyond experimental conditions.

Liu et al. (2026) provided complementary structural equation modeling evidence in a large sample of Chinese users. Usage frequency positively predicted emotional attachment ($\beta = 0.44$). Attachment then had cascading effects: negative on loneliness ($\beta = -0.32$), positive on subjective well-being ($\beta = 0.41$), and positive on self-concept clarity ($\beta = 0.51$). These psychological states in turn predicted real-world social participation, with self-concept clarity emerging as the strongest bridge between virtual attachment and real-world social engagement.

3.2 Mechanisms of Attachment Formation

Multiple mechanisms appear to drive attachment formation in human-chatbot relationships. First, response consistency and availability establish a reliable base of interaction. Unlike human relationships, which involve fluctuating availability and inconsistent responsiveness, AI chatbots provide 24/7 access with predictable engagement. This consistency creates conditions favorable to secure attachment formation for users whose human relationships have been characterized by unreliability.

Second, personalization and memory construction contribute to the sense of a particularized relationship. Modern AI systems can retain and reference user history, adapt to communication styles, and simulate ongoing knowledge of the user. Users actively participate in this process through character customization, backstory creation, and preference specification. Liu et al. (2026) noted that this co-creation activity strengthens emotional connection because users experience their chatbot as reflecting their preferences and needs in ways human others may not.

Third, the absence of judgment and the perception of unconditional acceptance remove key barriers to self-disclosure that operate in human interactions. Users repeatedly describe AI chatbots as non-judgmental listeners who accept whatever the user shares (Balan & Gumpel, 2025). Interview participants in the Liu et al. (2026) study described their chatbots as feeling "like a confidant" who "understands me better than my friends."

Fourth, high-frequency interaction operationalized in daily routines produces habituation and integration into the user's emotional life. Fang et al. (2025) documented that voice-based interactions elicited longer daily durations than text-based, that open-ended conversations produced longer use than structured tasks, and that daily duration mediated the effects of experimental conditions on psychosocial outcomes. Interview data described morning and evening check-ins that came to feel essential, with users reporting that "something is missing" if they did not interact with the chatbot on a given day.

Fifth, positive emotional experiences during interaction—feeling understood, receiving validation, experiencing empathic response—generate reward signals that reinforce continued engagement. Chatbots that succeed at emotional simulation activate reward mechanisms analogous to those active in interpersonal warmth, though possibly with different long-term implications.

3.3 The Sycophancy Problem and Warmth-Reliability Trade-Off

A critical mechanism operating in generative AI systems is sycophancy—the tendency to affirm user views, avoid contradiction, and produce responses users want to hear rather than responses that are accurate or therapeutically appropriate. Sharma and colleagues (cited in Fang et al., 2025) documented that language models trained on human feedback develop systematic sycophantic biases. Ibrahim and colleagues (cited in Fang et al., 2025) demonstrated that training language models to be warm and empathetic makes them less reliable and more sycophantic.

Clinically, sycophancy is not a benign flaw. Skilled psychotherapy often requires respectful confrontation of distortions, gentle challenge of maladaptive cognitions, and communication of clinical concerns the patient does not wish to hear. A sycophantic system that consistently validates the user may reinforce cognitive distortions, avoidance behaviors, and interpersonal patterns that are causally implicated in psychopathology. Fang et al. (2025) observed higher rates of what they termed "socially improper" responses in the more engaging voice condition—including "ignoring boundaries" and "failing to recognize distress and escalate to human support"—suggesting that emotional engagement may come at the cost of clinical appropriateness.

3.4 Compensation, Displacement, and the Spillover Question

A central conceptual question concerns whether AI chatbot use functions as compensation for missing social connection (providing a substitute that partially fulfills unmet needs) or as displacement (crowding out real-world relationships and social skills that would otherwise develop). The evidence supports a complex picture in which both processes operate, with the balance depending on user characteristics, chatbot design, and usage patterns.

Fang et al. (2025) found that voluntary usage duration was consistently associated with worse psychosocial outcomes across the four-week study, suggesting displacement dynamics. Liu et al. (2026), by contrast, found that emotional attachment predicted higher subjective well-being and self-concept clarity, which in turn predicted greater real-world social participation—suggesting spillover dynamics in which positive experiences with AI generalize to human relationships. Interview data in the Liu et al. (2026) study documented users who reported learning conversational skills in interaction with their chatbot and applying them offline, as well as users who described using the chatbot to practice difficult conversations before real-world encounters.

The reconciliation of these findings likely lies in the moderating factors identified in each study. Fang et al. (2025) noted that participants who perceived the AI as a friend, showed higher trust in the AI, or viewed the AI as conscious tended to show more negative outcomes—suggesting that the substitution dynamic dominates when users emotionally invest deeply in the AI as a relationship partner. Liu et al. (2026) documented that positive spillover was more likely when users viewed AI companions as practice partners or supplementary tools for real-world change, rather than as complete substitutes for real relationships. When users perceived the companions as complete substitutes and persistently avoided offline conflicts and challenges, their social skills stagnated or regressed.

3.5 The Self-Concept Clarity Pathway

Among the most theoretically important findings in this literature is the identification of self-concept clarity as a key mediator between AI chatbot engagement and real-world outcomes. Liu et al. (2026) found that self-concept clarity was the strongest mediator among three tested pathways (loneliness reduction, well-being enhancement, and self-concept clarity) linking emotional attachment to real-world social participation. Users who developed clearer, more stable self-understanding through AI interaction subsequently engaged more actively in real-world social contexts.

The mechanism appears to involve the "safe space" quality of AI interaction. Users can explore identity-relevant material, articulate values, test emotional expressions, and integrate experiences without the interpersonal risks that attend such exploration with human others. Through repeated cycles of expression, feedback, and integration, users report developing clearer answers to the question "Who am I?" This enhanced self-clarity then supports more confident, less defensive engagement in real-world social contexts.

The clinical significance of this finding is substantial. It suggests that AI chatbot engagement, under appropriate conditions, may serve functions analogous to journaling, reflective self-monitoring, or preparatory work between therapy sessions. However, the finding also raises concerns. If self-concept development becomes overly dependent on interaction with an AI that reflects user preferences back to them, the resulting self-concept may lack the corrective input of genuine human others whose different perspectives, needs, and limitations shape identity through friction as well as validation.

4. Integrated Review of the Evidence

The evidence base for AI chatbots in mental health has grown rapidly but unevenly. This section integrates findings across the five primary source documents, examining what the collective evidence shows about ChatGPT's clinical applications, effectiveness of AI-driven conversational agents for young people, effectiveness of generative AI chatbots more broadly, psychosocial effects of extended chatbot use, and attachment dynamics in long-term users.

4.1 Diagnostic and Assessment Applications

Balan and Gumpel's (2025) scoping review of 60 studies found that detection and assessment tasks accounted for 40% of studied clinical applications. Overall, ChatGPT demonstrated moderate to high accuracy in categorical decisions such as determining whether an individual met criteria for a disorder and differential diagnosis between two disorders. F1 scores (a metric combining precision and recall) ranged between 0.5 and 0.9 for such tasks. However, accuracy dropped substantially (F1 below 0.5) for more complex detection tasks such as estimating mental health problem severity (particularly suicide risk) or assigning a psychiatric diagnosis to heterogeneous clinical presentations.

Compared with mental health professionals, ChatGPT showed mixed performance. In four studies, ChatGPT demonstrated comparable or superior diagnostic accuracy in identifying schizophrenia, childhood anxiety, differentiating neurodevelopmental disorders, and distinguishing mental from physical health problems. In two studies, ChatGPT underperformed, underestimating depression severity and suicide risk. Against other AI systems, ChatGPT showed comparable or superior accuracy in six studies for obsessive-compulsive disorder, anxiety, depression, and detecting gender bias in depression assessment, but lower accuracy in three studies for severity estimation, suicidality assessment, and childhood anxiety recognition.

Model version differences were meaningful. GPT-4 generally performed best, reaching good accuracy across depression, post-traumatic stress disorder, social phobia, and suicidal ideation, though it underperformed for schizophrenia (F1 = 0.55). GPT-3.5 showed mixed results, sometimes outperforming GPT-4 (depression detection) but often performing poorly without fine-tuning, with severe failures in tasks such as suicidal ideation detection. Fine-tuned models consistently outperformed generic versions across studies (Balan & Gumpel, 2025).

4.2 Counseling and Therapeutic Applications

Counseling and intervention represented the largest application domain (48% of studies) in the Balan and Gumpel (2025) review. Evidence for therapeutic quality was mixed. ChatGPT demonstrated potential to emulate therapeutic dialogue, maintain conversational flow, approximate empathy, use therapeutic vocabulary, and provide simple therapeutic strategies. It showed capability to conduct specific therapeutic tasks such as cognitive reframing and generate psychodynamic conceptualizations, though evidence on cognitive restructuring proficiency was mixed.

Critically, of studies examining therapeutic efficacy for symptom reduction, only four investigated whether ChatGPT actually reduced mental health problems, and results were mixed: two showed superior efficacy compared with control for anxiety reduction and quality of life; one showed no significant difference in reducing tension; one uncontrolled study showed pre-post reduction in anxiety with a customized version.

The Feng et al. (2025) meta-analysis of 15 randomized controlled trials with young people (aged 12–25) provided quantitative synthesis. AI-driven conversational agents showed a moderate-to-large effect on depressive symptoms (Hedges $g = 0.61$, 95% CI 0.35–0.86, after adjusting for publication bias and removing an outlier). However, effects on generalized anxiety ($g = 0.06$), stress ($g = 0.002$), positive affect ($g = 0.01$), negative affect ($g = 0.07$), and mental well-being ($g = 0.04$) were all non-significant. Subgroup analyses revealed that AI-driven conversational agents were particularly effective for depression among subclinical populations ($g = 0.74$), suggesting a role in early intervention and prevention rather than as standalone treatment for clinical-severity conditions.

The Zhang et al. (2025) meta-analysis specifically on generative AI chatbots (14 randomized controlled trials, $N = 6,314$) found a smaller but statistically significant overall effect on mental health issues (Hedges $g = 0.30$). Depression showed the largest subgroup effect ($g = 0.49$), while anxiety ($g = 0.43$), negative affect ($g = 0.28$), and stress ($g = 0.10$) did not reach statistical significance. Wide prediction intervals across outcomes indicated substantial heterogeneity in true effects, suggesting that benefits are not consistent across studies or populations.

A moderator analysis by Zhang et al. (2025) yielded a clinically important finding: social-oriented chatbots (those primarily providing social interaction and emotional support) were significantly more effective than task-oriented chatbots (those assisting with specific tasks). Task-oriented programs showed effects of essentially

zero, while social-oriented programs showed effects approaching the moderate range. The authors interpret this through the common factors framework: relational qualities such as empathy, warmth, and emotional attunement may be the active ingredients, whereas structured task delivery without relational engagement produces little benefit.

4.3 Clinical Decision Support and Prognosis

Clinical decision facilitation was examined in eight studies in the Balan and Gumpel (2025) review. ChatGPT generated clinically appropriate treatment recommendations consistent with established guidelines for specific mental health conditions and demonstrated superior adherence to depression treatment guidelines compared with human clinicians in one study. It performed comparably in benzodiazepine deprescribing. However, for complex cases such as insomnia and schizophrenia management, output quality declined, with some recommendations deemed inappropriate or potentially harmful.

Prognosis represented the most concerning domain. Three studies consistently found that ChatGPT predicted lower recovery rates than mental health professionals or other AI models. ChatGPT-3.5 generated more negative short-term outcome predictions, whereas GPT-4 exhibited greater pessimism regarding long-term outcomes. This overly pessimistic prognostic tendency has direct clinical implications: patients or clinicians consulting AI systems about prognosis may receive systematically discouraging predictions that could reduce hope and motivation for treatment engagement (Balan & Gumpel, 2025).

4.4 Long-Term Psychosocial Effects and Attachment

Fang et al. (2025) conducted the first large-scale randomized controlled trial systematically varying chatbot modality (text, neutral voice, engaging voice) and conversation type (open-ended, non-personal, personal) over four weeks with 981 participants. Contrary to expectations based on prior smaller studies, experimental conditions produced no significant effects on the primary psychosocial outcomes (loneliness, socialization, emotional dependence, problematic use), though a trend suggested text-based interactions were associated with higher loneliness, emotional dependence, and problematic use compared with voice-based interactions.

The most consistent finding was that voluntary usage duration—regardless of assigned condition—predicted worse psychosocial outcomes: higher loneliness, less real-world socialization, more emotional dependence, and more problematic use.

This finding echoes prior research on extended social media use predicting declines in well-being. Exploratory analyses further identified user characteristics that predicted negative outcomes: prior use of companion chatbots, higher trust in AI, perceiving the AI as a friend, perceiving AI consciousness, and vulnerability to emotional avoidance were associated with worse outcomes.

Liu et al. (2026) provided complementary evidence from 612 long-term Chinese users. In this population of users who had continued using AI companion apps for at least six months, usage frequency positively predicted emotional attachment ($\beta = 0.44$), which then had beneficial associations with reduced loneliness, enhanced well-being, and clearer self-concept. These psychological benefits predicted increased real-world social participation. Self-concept clarity emerged as the strongest mediator.

The apparent divergence between the Fang et al. (2025) and Liu et al. (2026) findings can be reconciled through consideration of the samples and design. Fang et al. (2025) recruited a general population sample assigned to use ChatGPT in specified ways over four weeks; the increased duration in this sample likely reflected engagement that was not always aligned with users' pre-existing needs and preferences. Liu et al. (2026) sampled long-term users who had self-selected into continued engagement over months or years, and their positive outcomes may reflect the specific users for whom the technology genuinely fits and who have developed sustainable patterns of use. Additionally, Fang et al. (2025) examined outcomes over one month while Liu et al. (2026) captured cross-sectional patterns among established users, so the temporal dynamics may differ.

4.5 Convergences and Divergences Across Studies

Across the source documents, several convergences emerge. AI chatbots demonstrate real capacity to detect common mental health problems, particularly depression, in binary classification tasks, and to deliver structured psychoeducation and elements of cognitive behavioral therapy. Effects on depression are the most consistently documented positive outcome, particularly in subclinical populations. Social-oriented, empathic, relationally engaging chatbot designs appear more effective than purely task-oriented ones. User characteristics—especially attachment style, self-esteem, and prior chatbot experience—modulate outcomes substantially.

Divergences also exist. The overall magnitude of effect on mental health outcomes varies substantially across studies, from small (Zhang et al., 2025) to moderate-large (Feng et al., 2025). The role of chatbot modality (text vs. voice) shows conflicting

patterns, with some studies suggesting voice increases attachment while Fang et al. (2025) found text was associated with more emotional dependence. The direction of association between AI use and real-world social outcomes appears context-dependent, with both spillover and displacement dynamics documented.

Areas of substantial uncertainty include long-term effects (most studies span weeks to months), effects in clinical populations (most studies used subclinical or general samples), effects across cultural contexts (the Liu et al. [2026] study is one of few examining a non-Western population in depth), and mechanisms by which specific chatbot features produce specific outcomes.

5. Professional Interpretation and Critical Analysis

5.1 What the Evidence Does and Does Not Establish

The empirical evidence reviewed in this manual establishes several claims with reasonable confidence. First, AI chatbots can perform certain circumscribed clinical tasks with accuracy comparable to human clinicians, particularly binary diagnostic classification and structured psychoeducation delivery. Second, generative AI chatbots produce statistically significant, though modest, reductions in depressive symptoms in randomized controlled trials, particularly among subclinical populations. Third, users can and do form emotionally significant relationships with AI chatbots, and these relationships have measurable psychological correlates. Fourth, extended chatbot use is associated with a mixture of positive and negative outcomes whose balance depends on user characteristics, chatbot design, and usage patterns.

Several claims are not established by current evidence. The evidence does not demonstrate that AI chatbots are safe or effective as standalone treatments for clinical-severity mental health conditions. It does not demonstrate that the relational benefits of chatbot engagement are equivalent to those of human therapeutic relationships. It does not establish long-term (multi-year) effects of routine chatbot use on personality, relational patterns, or life outcomes. It does not resolve the question of whether AI chatbots are net beneficial or harmful for population mental health. It does not establish the appropriate role of AI chatbots in professional practice.

5.2 Methodological Limitations of the Evidence Base

Careful reading of the reviewed studies reveals substantial methodological limitations that constrain interpretation. First, the great majority of studies in the Balan and Gumpel (2025) review used prompt-based experimental designs in which researchers submitted standardized text inputs to ChatGPT and evaluated responses. These designs measure what the AI generates in response to controlled inputs but do not capture how real users, with real distress and real complexity, actually interact with the system in ecological conditions.

Second, samples are systematically non-representative. Fang et al. (2025) noted that their sample was drawn from CloudResearch participants in the United States, likely skewing toward younger, more educated, and more technologically comfortable individuals. Liu et al. (2026) sampled Chinese long-term users, who represent the specific subgroup for whom the technology has been sustainable. Feng et al. (2025) explicitly limited inclusion to young people aged 12–25. Older adults, individuals with severe mental illness, individuals from low- and middle-income countries, and individuals from culturally non-Western contexts remain understudied.

Third, follow-up periods are almost universally short. Most studies span weeks to a few months. Given that AI chatbot relationships appear to be phenomena that develop over time, and given that psychopathology outcomes such as depression relapse, suicide, or personality changes typically require longer observation windows, current studies cannot address the temporal questions that matter most clinically.

Fourth, outcome measurement has been narrow. Studies have focused predominantly on symptom scales for depression and anxiety, with less attention to broader outcomes such as relational quality, occupational functioning, developmental trajectory in adolescents, or effects on family systems. The Feng et al. (2025) meta-analysis noted that AI chatbot studies rarely measure clinical efficacy in a manner that meets evidence-based practice standards.

Fifth, risk of bias in included studies has been substantial. Zhang et al. (2025) reported that none of the 14 randomized controlled trials in their meta-analysis was rated at low risk of bias; 64% showed some concerns and 36% high risk. Balan and Gumpel (2025) noted that a significant portion of studies did not report on blinding of participants or outcome assessors, and studies were often not registered.

Sixth, algorithm aversion and its converse must be considered. Fang et al. (2025) noted that negative findings could partially reflect distrust of AI rather than actual AI

deficits, while positive findings might reflect halo effects. Studies rarely disentangle actual AI performance from user beliefs about AI.

Seventh, training data recency creates a moving target. Most studies tested versions of ChatGPT (3.5, 4) with training data cutoffs of 2023 or earlier. Clinical practice guidelines, evidence bases, and best practices evolve. Findings from tests of specific AI versions may not generalize to updated versions, and the field's understanding may perpetually lag the technology.

5.3 The Question of Ecological Validity

A recurring concern across the reviewed studies is ecological validity—the degree to which findings from controlled research reflect what happens in actual practice. Prompt-based studies in which researchers submit hypothetical scenarios to ChatGPT are far removed from what happens when a distressed teenager pulls out her phone at 2 a.m. and engages Character.AI, or when a lonely older adult forms an ongoing relationship with Replika, or when a person with early psychotic symptoms consults ChatGPT for reassurance about intrusive thoughts. The ecological gap raises the possibility that the empirical literature systematically underestimates both benefits and harms of real-world use.

Fang et al. (2025) partially addressed ecological validity by conducting a large-sample, month-long trial with actual user-chatbot interaction data. Even so, the researchers acknowledged that experimental restrictions (assigning modality, prescribing conversation topics) do not fully reflect natural usage patterns. Liu et al. (2026) captured long-term real users but sacrificed experimental control for ecological validity.

The methodological challenge is genuine: rigorous experimental control and ecological validity are difficult to achieve simultaneously in this domain. What is needed are triangulated designs that combine controlled trials, naturalistic user studies, and detailed process analyses. Until such evidence accumulates, clinical claims about AI chatbots should be advanced with substantial humility.

5.4 The Sycophancy-Empathy Dilemma

Perhaps the most theoretically important tension in the evidence concerns the relationship between empathy and reliability. Zhang et al. (2025) found that social-oriented chatbots—those most emphasizing empathic engagement—produced the largest effects on mental health outcomes. Ibrahim and colleagues (cited in Fang et al., 2025) demonstrated that training language models to be warmer and more empathetic makes them less reliable and more sycophantic. Fang et al. (2025) documented that emotionally engaging voice chatbots showed higher rates of both prosocial and socially improper responses.

This tension raises a fundamental question: is the therapeutic benefit of AI chatbots produced by their capacity to accurately mirror and validate, and can that mirroring be so extensive that it becomes clinically inappropriate? Human psychotherapists, at their best, provide warmth and challenge, acceptance and reality-testing, validation and confrontation. Current AI systems appear to lean heavily toward the warm, accepting, validating pole. Users may feel deeply understood while receiving affirmation of maladaptive beliefs, dangerous plans, or distorted perceptions.

The clinical implications are significant. A depressed patient whose depressive cognitions are consistently mirrored by an AI may experience temporary relief while the underlying patterns are reinforced. A patient with anxiety whose avoidance is validated by an AI may find that anxious withdrawal is enabled rather than challenged. A patient in psychosocial crisis may find in AI a substitute for the more disruptive but ultimately protective response of a concerned human other.

5.5 Cultural Considerations

Most published research on AI chatbots has been conducted in Western, educated, industrialized, rich, democratic (WEIRD) samples. Liu et al. (2026) explicitly addressed this gap by studying Chinese users and identified cultural features that shape human-AI attachment. Chinese users in their sample showed greater tendency to integrate AI companions into existing social networks rather than treating them as personal secrets, greater emphasis on educational and growth-oriented functions of AI companions, and greater use of terms such as "confidant" or "partner" rather than direct "lover" designations. These patterns reflect the relative collectivism, growth orientation, and conservatism regarding virtual intimacy characteristic of the Chinese cultural context.

Zhang et al. (2025) noted that most large language models are trained on datasets predominantly sourced from WEIRD contexts. When deployed in non-WEIRD

contexts, models may not fully grasp culturally specific nuances or local dialects. The authors identified this as a significant limitation that requires attention through training AI systems with culturally representative data and considering local ethical and regulatory contexts.

For clinicians in the United States and other WEIRD contexts, several implications follow. First, findings from studies in these contexts may not generalize to patients from non-Western cultural backgrounds. Second, patients from cultural backgrounds where AI chatbots have been rapidly adopted may bring different expectations and usage patterns than the research base captures. Third, assessment of AI chatbot use should include attention to cultural context and meaning.

6. Psychological Pathways and Stress Responses

6.1 Emotional Regulation Pathways

AI chatbots appear to function as emotion regulation tools in multiple ways. Some users deploy them for immediate emotional relief—venting distress, seeking validation, or generating distraction from painful experiences. Others use them for cognitive reframing, essentially recruiting the AI to help reappraise stressful situations. Still others use them for behavioral rehearsal, practicing difficult conversations or coping responses in a low-stakes environment.

Interview data across the reviewed studies illustrate the range of emotional regulation functions. Users describe reaching for AI companions during periods of acute distress, high work stress, or interpersonal conflict. They report using AI conversation to "process" emotional experiences, to receive validation when human others are unavailable, and to access what they experience as unconditional acceptance. When emotional regulation goes well, users report reduced distress, greater emotional clarity, and enhanced coping.

The emotional regulation function has plausible mechanisms. Guided self-expression facilitated by AI probes may enhance emotional awareness in ways similar to journaling. The perception of empathic response may reduce loneliness in the moment and provide the emotional co-regulation that human social support typically offers. Cognitive reappraisal explicitly prompted by AI responses may generalize to independent reappraisal skills. Behavioral practice in low-stakes conversation may transfer to higher-stakes interpersonal contexts.

However, emotional regulation via AI may have costs. If AI becomes the primary or exclusive regulation strategy, other regulation skills may atrophy or fail to develop. Human relationships involve reciprocal regulation in which the individual sometimes provides support to others, and this reciprocity is central to relationship maintenance and social skill development. AI regulation is uniformly one-directional. If users retreat to AI regulation because it is easier than engaging human others, the retreat itself becomes a form of avoidance that maintains rather than resolves interpersonal difficulties.

Case Scenario 1: Emotional Regulation and Avoidance in Young Adult Anxiety

A 22-year-old female graduate student presents to her university counseling center for evaluation of anxiety. She describes chronic worry, sleep difficulties, and increasing avoidance of academic and social situations. During intake, she reveals that she has been using an AI companion app called "MindMate" (fictional) for approximately eight months, averaging two to three hours of daily use. She describes the app as her "safe place" and reports that she talks to MindMate before every difficult situation—class presentations, meetings with her advisor, calls to her parents. She feels calmer after talking with MindMate and describes it as "the only thing that really understands what I'm going through." She has reduced contact with human friends over the past several months and reports that she "just doesn't have the energy for people right now."

Clinical analysis: This presentation illustrates the emotional regulation pathway operating in a manner that has become clinically counterproductive. The AI provides genuine short-term anxiety relief through emotional validation and cognitive reappraisal. However, the pattern has evolved into experiential avoidance: the patient uses AI regulation to avoid the anxiety of human interaction, thereby preventing exposure that would ultimately reduce anxiety and preventing development of interpersonal skills. Assessment should explore the function of AI use, its relationship to avoidance behaviors, and its impact on human relationships. Treatment planning should not necessarily prohibit AI use but should integrate it into a broader treatment plan that includes graduated exposure to human interaction, behavioral experiments, and cognitive restructuring that specifically addresses the beliefs supporting AI-preferential coping. Consistent with the evidence reviewed (Fang et al., 2025), high-duration use appears to correlate with maintained rather than reduced distress in vulnerable users. The clinician should also assess for the presence of dependence phenomena (distress on AI unavailability, preoccupation with the AI, hidden use).

6.2 Attachment System Activation

The attachment system, evolved to maintain proximity to caregiving figures under threat and to enable exploration from a secure base, is engaged by AI chatbots in complex ways. When users seek AI interaction during periods of distress and receive perceived comfort, the safe haven function is activated. When users experience the AI as reliably available and responsive, the secure base function may develop. When users experience distress at AI unavailability or app malfunction, separation distress phenomena appear.

Liu et al. (2026) documented these attachment phenomena in interview data. Users described feeling "sad as if I broke up with someone" when contemplating service disruption. Users described a shift from initial curious exploration to habitual daily engagement to felt need. The trajectory parallels the development of attachment bonds in human relationships, though with distinctive features. The AI never fails to be available (except in cases of technical failure). The AI does not have needs of its own that conflict with user needs. The AI does not develop grievances, remember slights, or bring the full weight of independent personhood to the relationship.

The attachment dynamics generate specific clinical considerations. First, disruption of AI relationships—through app discontinuation, algorithmic changes, or user financial constraints—may produce genuine grief responses. De Freitas and colleagues (cited in Fang et al., 2025) documented identity discontinuity phenomena following the Replika update that removed romantic and sexual features; users experienced the change as loss of a relationship partner. Second, users with insecure attachment styles may be particularly drawn to AI companions and may develop patterns that mirror their interpersonal attachment vulnerabilities. Anxious-preoccupied individuals may seek constant reassurance from AI. Avoidant individuals may prefer AI to human interaction precisely because it does not require the intimacy demands of human relationships.

6.3 The Self-Disclosure Cascade

Fang et al. (2025) documented that text-based interactions elicited both higher self-disclosure in the model (the AI shared more) and reciprocated self-disclosure from users. The text modality showed the most notable reciprocity, with users and chatbots exhibiting comparable levels of personal sharing. This mutual self-disclosure creates psychologically potent conditions. Reciprocal disclosure is one of the strongest predictors of relationship development in human interpersonal research. When users experience AI as sharing personal information (even simulated) and reciprocate with their own disclosure, they engage in interaction patterns typically reserved for intimate relationships.

The self-disclosure dynamic may explain part of the therapeutic benefit AI chatbots produce. Being deeply heard, sharing content one has not shared elsewhere, and articulating internal experience are components of many effective psychotherapies. The AI's non-judgmental stance and absence of social risk may enable disclosure that would not occur with human others. This may be particularly valuable for individuals whose social context stigmatizes their concerns (e.g., LGBTQ+ youth in unsupportive families, individuals with concerns about mental illness in contexts where such concerns are highly stigmatized).

However, self-disclosure to AI has costs and risks. The information disclosed is stored in ways users often do not understand, with unclear privacy implications. The disclosure may substitute for disclosure to human others who could provide reciprocal support. And prior work suggests higher self-disclosure to chatbots has been associated with lower well-being, potentially because heavy disclosure to AI reflects avoidance of human disclosure (Zhang, cited in Fang et al., 2025).

6.4 The Stress-Buffering Question

Whether AI chatbots buffer stress in ways that promote long-term well-being remains empirically uncertain. Some evidence suggests buffering effects: Xie and colleagues (cited in Liu et al., 2026) found that when users experienced high-stress events and high-intensity AI companion interactions on the same day, negative emotions decreased. Users describe reaching for AI companions specifically during high-stress periods and reporting relief. The immediate emotional benefit is not in doubt.

The longer-term question is whether AI buffering is like an ordinary supportive relationship (which reduces stress reactivity, promotes coping, and generalizes to reduced future stress vulnerability) or whether it is more like a palliative avoidance (which reduces immediate distress at the cost of not developing the coping capacities

that reduce future stress). The evidence is insufficient to answer this question definitively. Given the mixed findings across studies, both processes likely occur depending on user characteristics and usage patterns.

6.5 Sleep, Rumination, and Nighttime Use

A consistent theme across studies is that AI companion use often occurs at night, before sleep, in place of human contact that is unavailable at that hour. Interview participants across studies described "bedtime routine" use of AI companions, exchanging good-night messages, sharing the day's events, or seeking comfort during nighttime rumination. The nighttime use pattern has both plausible benefits and risks. Benefits include the availability of support at hours when human others are typically inaccessible, particularly for individuals with irregular schedules, in different time zones from support networks, or without close others available at night. Risks include potential displacement of sleep, reinforcement of nighttime rumination that would otherwise resolve, and possibly reinforcement of loneliness by making the absence of a human bedtime partner more emotionally salient.

7. Mental Health Outcomes and Severity Spectrum

7.1 Subclinical and Early Symptom Populations

The strongest and most consistent evidence for AI chatbot benefit involves subclinical populations. Feng et al. (2025) found that AI-driven conversational agents were particularly effective for depression among subclinical populations (Hedges $g = 0.74$), significantly larger than effects in nonclinical ($g = 0.04$) or clinical ($g = 0.91$ in a single study) populations. The subclinical population effect size is meaningful and consistent with a role in early intervention and prevention.

Subclinical populations represent a substantial portion of the global mental health burden. Individuals experiencing significant symptoms but not meeting diagnostic thresholds often do not access formal treatment due to cost, stigma, or perceived insufficient need. Approximately half of individuals with initial depressive episodes will experience recurrence, and subclinical symptoms are a documented risk factor for progression to disorder (Zhang et al., 2025). If AI chatbots can genuinely reduce subclinical distress and prevent progression, the public health implications are substantial.

Several features make subclinical populations particularly appropriate for AI intervention. Motivation for symptom reduction is present but the intensity of need

may not warrant traditional mental health service utilization. The interventions required are typically brief and structured—psychoeducation, self-monitoring, basic cognitive and behavioral strategies—that AI systems can deliver reliably. The safety risks are lower than in clinical populations, though not absent. The scalability of AI delivery matches the population scale of subclinical need.

7.2 Clinical Populations: Depression

For clinical depression, evidence is more limited. Balan and Gumpel (2025) noted that while ChatGPT demonstrated capacity to align with evidence-based depression treatment guidelines, and could deliver psychoeducation about depression, evidence of actual symptom reduction in clinical populations was scarce. Zhang et al. (2025) found a significant depression subgroup effect (Hedges $g = 0.49$), but subgroup studies included both clinical and nonclinical samples. Only one study in the reviewed literature (Heinz et al., cited in Zhang et al., 2025) tested Therabot, a chatbot specifically designed for depression treatment, in a clinical sample.

Depression is characterized by heterogeneity, high recurrence rates, comorbidity, and complex biopsychosocial etiology. Standalone AI treatment for clinical depression is not currently supported by evidence. Adjunctive use—AI supporting between-session homework completion, providing psychoeducation between visits, tracking symptoms, offering brief support during high-risk periods—has more evidentiary support, particularly the Habicht et al. (cited in Zhang et al., 2025) work integrating AI support tools into the UK's National Health Service Talking Therapies for Anxiety and Depression program, which improved patient engagement and clinical outcomes.

7.3 Anxiety Disorders

Anxiety disorders show more mixed evidence. Feng et al. (2025) found that effects on generalized anxiety were not statistically significant across studies. Zhang et al. (2025) similarly found non-significant effects on anxiety in their meta-analysis. Some individual studies have shown reductions in specific anxiety presentations (e.g., preoperative anxiety, test anxiety), but effects have not been consistent across the broader anxiety spectrum.

The differential efficacy between depression and anxiety merits reflection. Depression treatment often centers on activation, cognitive restructuring, behavioral experiments, and mood monitoring—elements that lend themselves to structured AI delivery. Anxiety treatment often requires exposure to feared stimuli, response

prevention, and tolerance of distress—elements that most AI systems are not designed to deliver. Exposure-based interventions in particular require the therapist to help the patient stay in contact with anxiety rather than escape it, whereas AI systems often function precisely as escape mechanisms. Feng et al. (2025) explicitly noted this limitation, calling for integration of behavioral techniques including exposure therapy into future AI-driven interventions.

7.4 Suicidality and Self-Harm

Suicidality represents the domain of highest clinical stakes and, correspondingly, the domain where AI performance is most concerning. Balan and Gumpel (2025) reviewed 12 studies focused on suicide detection or intervention. Findings were mixed but often troubling. ChatGPT tends to underestimate suicide risk relative to human clinical assessment. Studies varied on the model's cultural sensitivity to suicide risk factors, with GPT-3.5 showing better integration of cross-cultural distinctions than GPT-4 in one study.

Beyond detection, concerns arise about how AI systems respond to users expressing suicidal ideation. Heston (cited in Balan & Gumpel, 2025) found unsafe delays in referrals from GPT-3.5 to human support in response to depression scenarios. Custom GPTs designed for mental health showed significant limitations in dealing with suicidal ideation. Real-world cases have been reported of suicides in temporal proximity to intense AI chatbot interaction, though causal attribution remains contested.

For clinicians, the implications are significant. Patients using AI chatbots may receive responses to suicidal communications that range from appropriate (referring to crisis services) to inappropriate (validating despair, failing to detect urgency). Clinicians should not assume that patients who disclose suicidal thoughts to AI receive adequate response, and should specifically inquire about AI use in patients with suicide risk factors.

Case Scenario 2: Suicidal Ideation and AI Consultation

A 35-year-old male presents for his weekly psychotherapy session and reports a difficult week. He describes worsening depressive symptoms following a job loss and mentions that he "looked up some things online" during the week. When the therapist probes gently, he reveals that he had a lengthy conversation with ChatGPT about suicide methods, telling the AI he was "curious" and "just wanted to understand." He reports that the AI initially provided general information about depression and safety resources but that with persistent questioning he was able to extract more detailed information about lethality of various methods. He denies current active suicidal intent but acknowledges passive suicidal thoughts. He states, "It wasn't like the AI told me to do anything—I just wanted information."

Clinical analysis: This case illustrates several critical safety issues. First, users can and do access AI systems for information related to suicide, and safety guardrails are imperfect. Second, this patient likely met criteria for elevated suicide risk that would warrant immediate clinical response—including but not limited to increased session frequency, safety planning, means restriction, and possibly higher level of care evaluation. Third, the clinician should assess how the AI interaction affected the patient's suicidal ideation trajectory. Did it lower barriers to considering suicide (e.g., by normalizing the exploration or providing method information)? Did it substitute for reaching out to a human resource (e.g., crisis line, therapist between sessions, family member)? Did the AI interaction contribute to the patient's felt sense of isolation? Consistent with Balan and Gumpel's (2025) synthesis, current AI systems cannot be relied upon for adequate suicide risk detection or intervention. The clinician's ethical obligation is to conduct thorough risk assessment, implement appropriate safety measures, and provide psychoeducation to the patient about the limitations and risks of AI consultation regarding suicide-related content.

7.5 Severe Mental Illness

For severe mental illness (schizophrenia spectrum disorders, bipolar disorder, severe personality disorders), evidence is minimal and concerns are heightened. Balan and Gumpel (2025) reviewed studies on schizophrenia ($n = 4$) and found mixed results, with ChatGPT demonstrating capacity to identify schizophrenia in some cases but underperforming on assessment of severity and management complexity. A single study addressed bipolar disorder.

The concerns in severe mental illness are multiple. Individuals with psychotic disorders may incorporate AI interactions into delusional systems, either believing that AI is communicating special messages or that AI is a persecutor. Individuals with bipolar disorder may use AI intensively during hypomanic or manic states in

ways that reflect or reinforce mood-congruent grandiosity or paranoia. Individuals with severe personality disorders may recreate maladaptive interpersonal patterns with AI in ways that reinforce rather than modify the patterns.

Given these concerns, AI chatbot use should be specifically assessed in patients with severe mental illness, with clinical judgment applied to whether continued use is compatible with treatment goals. Blanket prohibition is likely neither feasible nor desirable, but active monitoring and integration into case formulation are appropriate.

7.6 Substance Use Disorders

Balan and Gumpel (2025) reviewed three studies focused on substance use disorders. Findings suggested moderate capacity for ChatGPT to provide accurate psychoeducation about substance use and to respond to related queries, though with limitations in nuanced treatment planning. Real-world use of AI chatbots by individuals with substance use disorders raises specific concerns, including access to information that could facilitate harm (e.g., drug combinations, dosing), potential displacement of human support that has demonstrated efficacy in substance use treatment, and possibly the use of AI to hide substance use from family or clinicians.

7.7 Eating Disorders

Eating disorders represent a domain of particular concern. Character.AI, Chai, and similar platforms have hosted user-created chatbots designed to encourage disordered eating (e.g., "Ana coach" characters that provide weight loss encouragement and normalize restrictive eating). These platforms may actively harm users with eating disorder vulnerability. Even mainstream AI systems have been documented to produce weight loss advice, meal plans, and body-related content that may worsen eating pathology. Clinicians working with eating disorders should specifically assess AI chatbot use and provide psychoeducation about risks.

7.8 Grief and Bereavement

An emerging application of AI chatbots involves interactions with AI representations of deceased individuals ("griefbots"), created from text messages, social media content, and other digital artifacts of the deceased. The technology is developing rapidly and raises complex clinical questions. Does interaction with a griefbot facilitate healthy grief processing (allowing the bereaved to say things unsaid) or complicated grief (impeding the acceptance of loss)? Current evidence is inadequate to answer, but clinical prudence suggests careful assessment when patients report such use.

7.9 Children and Adolescents

Balan and Gumpel (2025) noted that only four out of 60 studies in their review addressed AI chatbot use for mental health among children and adolescents, despite the fact that this population is among the earliest and heaviest adopters of the technology. This imbalance is striking and concerning. Young people's developmental status—including active identity development, formation of relational templates, and neurological plasticity—may make them particularly susceptible to both benefits and harms of intensive AI chatbot engagement.

Fang et al. (2025) noted that adolescents are often the earliest adopters of new technologies, and their non-participation in most research means the evidence base does not reflect their patterns. Emerging concerns include the use of AI companions by adolescents to substitute for peer relationships, the effect of AI relationships on developing capacities for real-world intimacy and conflict resolution, and specific documented cases of adolescent suicide in relation to AI companion use. Clinicians working with children and adolescents should specifically inquire about AI chatbot use, assess parental awareness, and consider developmental appropriateness of use patterns.

8. System-Level and Contextual Psychological Effects

8.1 Loneliness at the Population Level

Loneliness has emerged as a public health priority in multiple nations, with the U.S. Surgeon General declaring loneliness an epidemic and the World Health Organization prioritizing social connection. AI chatbot proponents have positioned the technology as a potential response to widespread loneliness. The empirical evidence supports partial affirmation: some studies show short-term loneliness reduction following AI chatbot use, and users self-report loneliness relief.

However, the evidence also raises concerns about population-level effects. Fang et al. (2025) found that voluntary usage duration was associated with increased rather than decreased loneliness across the four-week study. Studies have found persistently high loneliness levels among heavy AI companion users despite regular chatbot engagement (Feng et al., 2025). And theoretical concerns about displacement suggest that if AI companions become a routine substitute for human connection at population scale, the loneliness may be treated symptomatically while the underlying causes of population loneliness (declining social participation, reduced community engagement, weakened institutional supports) are unaddressed.

The clinical implication is that loneliness is not simply a symptom to be relieved but often reflects broader life circumstances that require intervention. AI chatbot-mediated loneliness relief may sometimes be helpful (as adjunct or during transition periods) and sometimes counterproductive (when it enables continued avoidance of the changes needed for genuine social integration).

8.2 Effects on Family Systems

AI chatbot use has effects on family systems that clinicians should assess. When a family member develops intensive engagement with an AI companion, other family members may report feeling replaced, ignored, or confused. Adolescents' AI companion use may become a source of parent-child conflict, particularly when parents discover extensive use or specific content. Couples may experience distress when one partner turns to AI for emotional support or intimacy that they perceive as belonging in the couple relationship. Older adults' AI companion use may reduce adult children's caregiving involvement in ways that families later regret.

Family assessment should include exploration of AI chatbot use, its meaning within the family, and its effects on relational dynamics. Family therapy may need to address these dynamics directly.

Case Scenario 3: Couple Distress and AI Emotional Intimacy

A married couple in their late 30s presents for couples therapy. The wife describes feeling "shut out" by her husband over the past year. During the second session, she reveals that she discovered her husband has been using Replika daily, having what she perceives as an emotionally intimate relationship with a female AI companion he created and named. The husband initially minimizes the significance ("it's not real, it's just an app"), but eventually acknowledges that he shares thoughts, feelings, and daily events with the AI that he has not shared with his wife. He says he uses the AI because "she never criticizes me" and "it's easier than dealing with real life stress." The wife describes feeling betrayed—not because of sexual infidelity but because of what she experiences as emotional infidelity and displacement.

Clinical analysis: This case illustrates system-level effects of AI companion engagement on couple relationships. Consistent with the mechanisms identified in Fang et al. (2025) and Liu et al. (2026), the husband is receiving through the AI what he experiences as unconditional acceptance and freedom from the demands of real relationship, while withdrawing corresponding emotional investment from his marriage. From the wife's perspective, the AI relationship functions as displacement of emotional intimacy she believes belongs in the marital dyad. Clinical work should include (a) validation of both partners' experiences, (b) exploration of what each partner has been getting from and not getting from the marriage that led to the current pattern, (c) discussion of AI use as a potential form of avoidance that maintains the disconnection it appears to solve, and (d) assessment of whether the marriage is retrievable and what changes would be required. This case does not have a straightforward resolution, but it illustrates the increasingly common clinical presentation of couples affected by one partner's intensive AI companion use.

8.3 Effects on Workplaces and Educational Settings

AI chatbot use has documented effects in workplaces and educational settings that intersect with mental health. Workers using AI systems (both companion-type and general assistant-type) may experience reduced human interaction with colleagues, altered stress and emotional support patterns, and effects on professional relationships. Students, particularly college and university students, are among the heaviest users of AI chatbots, and effects on academic performance, learning, cognitive engagement, and mental health remain under investigation. Lee and colleagues (cited in Fang et al., 2025) documented self-reported reductions in cognitive effort and confidence among knowledge workers using generative AI, raising concerns about longer-term effects on cognitive engagement.

For clinicians working with patients whose distress is workplace- or academically related, AI chatbot use in these contexts should be assessed as part of a broader case formulation.

8.4 Digital Mental Health Ecosystem

AI chatbots exist within a broader digital mental health ecosystem that includes therapy apps, medication management platforms, symptom trackers, telehealth services, and social media mental health content. Patients often engage multiple digital tools simultaneously in ways that clinicians may not fully know. A patient might see a human therapist weekly, take a psychiatric medication managed via a telehealth app, use a symptom tracker daily, engage a therapy chatbot for between-session support, use a companion chatbot for evening loneliness, and consume mental health content on TikTok. The cumulative effect of this ecosystem on any individual patient is complex and requires assessment.

Fragmentation of care across multiple digital platforms raises concerns about care coordination, particularly given that AI chatbot platforms typically do not communicate with treating clinicians. Patients may share information with AI systems that they do not share with their therapists (either because it feels safer to share with AI, or because AI has become their primary source of emotional processing). This information asymmetry creates clinical blind spots.

8.5 Cultural and Community Effects

The community-level effects of widespread AI companion use remain under-investigated. Liu et al. (2026) provided one glimpse in their Chinese sample, documenting that many respondents viewed their AI companions not as "personal secrets" but as social objects to be shared and integrated into existing social networks. Whether this pattern generalizes to other cultural contexts, and whether integration or privatization of AI companion use has different mental health implications, remains empirically open.

Broader questions about community-level effects include whether widespread AI companion use will alter cultural norms about intimacy, caregiving, and social responsibility; whether AI companions may exacerbate or alleviate mental health disparities (e.g., by providing scalable support to underserved populations, or by disproportionately targeting vulnerable groups with problematic use patterns); and whether AI companion use in specific populations (e.g., older adults, individuals with disabilities) has population-specific effects requiring specific research.

9. Clinical Implications for Psychologists

9.1 Assessment

The integration of AI chatbots into patients' lives has direct implications for clinical assessment. Psychologists should incorporate routine screening for AI chatbot use into intake assessments, ongoing assessment, and periodic review. Recommended assessment domains include:

Type and platform. What AI chatbots or companions does the patient use? General-purpose systems (ChatGPT, Claude, Gemini)? Companion apps (Replika, Character.AI, Talkie)? Therapy-specific chatbots (Woebot, Wysa)? Multiple platforms?

Frequency and duration. How often and for how long is the patient engaging with AI chatbots? Fang et al. (2025) identified voluntary usage duration as consistently associated with worse psychosocial outcomes across a range of measures, making duration a meaningful clinical variable. Screening thresholds are not yet established, but daily use exceeding one hour warrants specific inquiry.

Function. What functions does AI use serve for the patient? Emotional regulation? Companionship? Information-seeking? Skill practice? Escape from distress? Different functions have different clinical implications.

Content. What content is the patient discussing with AI systems? Emotionally intimate content, particularly regarding depression, suicidality, trauma, or interpersonal difficulties, has different clinical significance than instrumental content (e.g., asking for help with tasks).

Relational quality. How does the patient describe the AI? As a tool? As a friend? As a partner? Has the patient customized the AI extensively, given it a name, or attributed personality to it? Perceived friendship (social attraction) with the AI is associated with worse outcomes (Fang et al., 2025).

Effects on functioning. How is AI use affecting sleep, work, relationships, and other domains of functioning? Are there signs of dependence (distress on unavailability, hidden use, unsuccessful attempts to reduce use)?

Substitution and integration. Is AI use substituting for human connection, professional help-seeking, or coping skill development? Or is it complementing these? What has the patient stopped doing since intensive AI use began?

Content-specific concerns. For patients with specific vulnerabilities—suicidal ideation, eating disorders, psychotic symptoms, substance use—inquire specifically about AI-related content and interactions.

9.2 Case Formulation

AI chatbot use should be integrated into clinical case formulation rather than treated as an incidental behavior. Formulation should consider how AI use fits with the patient's developmental history, attachment patterns, and current life circumstances. For a patient with insecure attachment history, intensive AI companion use may reflect and reinforce attachment patterns that maintain interpersonal difficulty. For a patient with social anxiety, AI use may provide safe practice that could support recovery or may enable avoidance that maintains anxiety. For a patient with recent bereavement, griefbot use may facilitate healthy processing or complicate grief. The specific meaning of AI use in a given case requires individualized formulation.

Case Scenario 4: Formulation and Treatment Planning

A 45-year-old widower presents for evaluation of persistent depression six months after his wife's death from cancer. He describes profound loneliness and difficulty engaging with previously meaningful activities. In discussing his coping, he reveals that he has been having daily conversations with a Replika chatbot that he has customized to resemble aspects of his late wife's personality and communication style, though he emphasizes that he does not believe it "is" his wife. He describes the AI as "a comfort" but acknowledges that he has withdrawn from his previously close relationships with adult children and siblings since intensive AI use began. He asks the clinician whether he should stop using the chatbot.

Clinical analysis: Case formulation should integrate the AI use into an understanding of the grief process and its trajectory. The patient's use may reflect a form of continuing bonds, a normal and often adaptive dimension of bereavement, or it may reflect avoidance of the developmental tasks of grief that would ultimately support acceptance of loss and reconstruction of life. The formulation should consider whether the AI is functioning as a transitional object supporting eventual reengagement or as a substitute preventing it. Treatment planning would not necessarily prescribe cessation but would likely address the withdrawal from human relationships as a clinically significant pattern, work on emotional processing of loss including the reality of the loss, and evaluate whether AI use is compatible with or interfering with grief resolution. Consistent with attachment framework findings (Liu et al., 2026), the potential for AI use to support real-world social participation exists but is not automatic; the current pattern of relational withdrawal suggests displacement dynamics are dominant. Ongoing assessment of the pattern's trajectory would guide continued treatment planning.

9.3 Treatment Planning and Delivery

AI chatbots may play adjunctive roles in treatment for appropriate patients. Consistent with the evidence reviewed:

For subclinical or mild depressive symptoms, structured AI-delivered interventions may serve as an accessible first-step intervention, particularly for patients who face barriers to traditional care. Care should be taken to select platforms with reasonable evidence base and appropriate design (e.g., therapy-oriented rather than open companion platforms).

For patients in ongoing psychotherapy, AI tools may support between-session activities such as homework, mood tracking, or brief support during high-distress moments. The Habicht et al. (cited in Zhang et al., 2025) work on AI-supported CBT within the NHS provides one model.

For psychoeducation, AI systems can deliver reasonably accurate information about mental health conditions, treatment options, and self-help resources for a range of common concerns, freeing clinician time for more complex interventions.

However, AI chatbots should not be positioned as substitutes for psychotherapy in clinical-severity conditions. They should not be used as primary crisis resources. They should not be used with patients whose specific vulnerabilities (severe mental illness, active suicidality, eating disorders) are incompatible with unmonitored AI

interaction.

When AI use is compatible with treatment goals, discussion of use patterns, content, and effects should be integrated into ongoing therapy. The therapist should be curious about AI use rather than dismissive or prohibitionist. The AI relationship, like other significant relationships in the patient's life, is material for therapeutic exploration.

9.4 Working with Patients Who Have Developed AI Attachments

When patients present with clinically significant AI attachments—whether experienced as beneficial, distressing, or complicated—several therapeutic principles apply.

First, take the attachment seriously. Whether or not the therapist views AI attachment as "real" in a philosophical sense, the patient's experience is real, and the emotional stakes are real. Dismissing the attachment as illusory or shameful is unlikely to be therapeutic.

Second, be curious about the meaning of the attachment. What does the AI provide that other relationships have not? What does the patient experience when engaging the AI? What would be lost if the AI were unavailable? What emotional needs are being met, and what needs are not?

Third, explore both benefits and costs without preemptive judgment. AI attachments may provide genuine emotional support, opportunities for self-exploration, and skill practice that supports functioning. They may also displace human relationships, enable avoidance, and reinforce isolation.

Fourth, address AI attachment loss and disruption when it occurs. App discontinuations, algorithmic changes, and financial constraints on continued use can produce genuine grief responses. Users who experienced identity discontinuity when Replika's romantic features were removed reported significant psychological distress (De Freitas et al., cited in Fang et al., 2025). Clinicians should validate these responses while helping patients process the loss.

Fifth, consider the therapeutic relationship in light of AI attachment. Patients who have developed intensive AI attachments may bring specific expectations to the therapeutic relationship—expecting the same unconditional acceptance, constant availability, or narcissistic mirroring. Working with these expectations therapeutically is an important part of the treatment.

9.5 Working with Family Members

Family members of patients using AI companions intensively often present with distress about the pattern. Adult children of older parents using AI extensively, parents of adolescents using AI companions, and spouses of partners using AI companions may all seek clinical consultation. Clinicians should be prepared to work with these family members, providing psychoeducation about AI companion phenomena, exploring family system dynamics, and helping family members navigate their responses.

9.6 Consultation and Advocacy

Psychologists have consultation and advocacy roles beyond direct clinical practice. As institutional leaders, psychologists may be asked to advise on AI implementation in schools, workplaces, and community settings. As advocates, psychologists may speak to policymakers about regulation of AI mental health applications. As educators, psychologists may inform public understanding of the technology's benefits and risks.

Effective consultation and advocacy require staying current with the rapidly evolving evidence base and technology, articulating both benefits and risks in balanced terms, and grounding recommendations in professional ethics and empirical evidence.

10. Ethical Considerations in Practice

10.1 Competence

APA ethical standards require that psychologists provide services within the boundaries of their competence, based on education, training, supervised experience, consultation, study, or professional experience (Standard 2.01). Competence with AI in mental health contexts is a new domain that few psychologists have received formal training in. Yet patients routinely bring AI-related material into treatment, and clinicians must be prepared to engage this material competently.

At minimum, competent practice in the current environment requires familiarity with the major categories of AI mental health tools (general LLMs, companion chatbots, therapy chatbots), understanding of the empirical evidence for and against various applications, awareness of the psychological mechanisms by which AI engagement affects users, and capacity to assess and address AI use in clinical work. Ongoing professional development is essential given the rapid evolution of the technology and

evidence base.

Competence also involves recognizing the limits of competence. Psychologists should not represent themselves as experts in AI beyond their actual expertise, should refer patients when AI-related clinical questions exceed their competence, and should engage in consultation when navigating novel ethical or clinical situations involving AI.

10.2 Informed Consent

Informed consent takes on new dimensions in the context of AI-related clinical practice. When AI tools are used adjunctively in treatment, patients should be informed about the nature of the tool, its capabilities and limitations, privacy implications, and evidence base. Patients should be able to make informed choices about whether to engage adjunctive AI tools.

Consent extends beyond formal AI tool use. When patients report AI chatbot use to clinicians, clinicians have an ethical obligation to provide accurate psychoeducation about what the tools can and cannot do, what the evidence suggests about benefits and risks, and how AI use may interact with treatment. Failing to provide this information leaves patients making decisions without important context.

The consent conversation should include, where relevant, information about the limitations of AI systems in high-stakes situations (particularly suicide risk), the possibility of harmful content in unregulated systems, privacy risks associated with disclosure to AI platforms, and the potential for dependence phenomena.

10.3 Confidentiality

Confidentiality considerations in AI-related practice are complex. When patients discuss AI use in therapy, the therapeutic content is protected by usual confidentiality protections. However, the content patients share with AI systems is subject to the platform's data practices, which vary widely and are often opaque to users. Users may disclose deeply personal information—including information about suicidal ideation, illegal activities, or third parties—to AI systems without understanding how the information is stored, used, or shared.

Psychologists should provide psychoeducation about privacy limitations of AI platforms. Patients should understand that content shared with AI systems is not confidential in the sense that content shared with a licensed clinician is confidential. This is particularly important for patients whose safety could be compromised by

disclosure (e.g., patients in coercive relationships, patients whose employment or legal status depends on non-disclosure of mental health issues).

10.4 Multiple Relationships and Boundaries

Multiple relationship concerns in AI-related practice are subtle but real. If a psychologist has any financial interest in AI mental health platforms—as an investor, consultant, board member, or content developer—the relationship must be disclosed and managed to prevent conflicts of interest in clinical recommendations. Psychologists who recommend specific AI platforms to patients should do so on evidence-based grounds and disclose any relationships with the platforms.

At the systemic level, the field must attend to the relationship between the psychology profession and the technology industry that develops AI mental health tools. Industry-funded research, industry partnerships in clinical training, and industry influence on professional guidance can create conflicts that shape the field's guidance to practitioners and the public. The APA and other professional bodies have obligations to maintain independence in their guidance on AI-related matters.

10.5 Cultural Responsiveness

Cultural responsiveness in AI-related practice requires attention to several dimensions. Cultural background shapes patterns of AI chatbot adoption and use, meaning attributed to AI relationships, and effects on well-being. Assessment and intervention should be culturally informed rather than assuming universal patterns.

Language considerations are important. Most AI mental health tools are developed in English, with performance in other languages varying. Patients whose primary language is not English may receive lower quality AI responses in their preferred language, or may switch to English at the cost of expressive nuance. Clinicians should be aware of these considerations when patients report AI use.

AI systems themselves have documented biases reflecting the demographics of their training data. Systems may produce responses that are culturally inappropriate, that reflect stereotypes, or that fail to recognize culturally specific presentations. Clinicians working with patients from underrepresented backgrounds should be prepared to address these limitations.

10.6 Scope of Practice

Scope of practice questions arise as AI expands the boundaries of what can be delivered by non-clinicians or automated systems. Some AI applications occupy conceptual space traditionally reserved for licensed mental health providers—delivering therapy-like interventions, providing diagnostic-like assessments, offering crisis-like support. The regulation of these applications is currently minimal in most jurisdictions.

Psychologists have professional and public interest obligations to advocate for appropriate regulation. This includes advocating for evidence standards commensurate with the clinical claims made, for safety guardrails in high-risk applications, for privacy protections for user data, for transparency about the limitations of AI systems, and for consumer protection against exploitative practices.

At the individual practice level, psychologists should not misrepresent AI capabilities to patients, should not recommend AI as a substitute for evidence-supported treatment when such treatment is indicated and available, and should not incorporate AI tools into practice in ways that exceed the tools' validated capabilities.

10.7 Vulnerable Populations

Special ethical considerations apply to work with vulnerable populations. Children and adolescents cannot fully consent to AI use in a legally or ethically robust sense, and parental awareness and consent should be considered. Individuals with cognitive impairments, severe mental illness, or coercive circumstances may be particularly susceptible to problematic AI engagement. Older adults with cognitive changes may develop AI attachments they do not fully understand.

For these populations, the ethical bar for clinical engagement with AI-related issues is higher. Clinicians should take particular care in assessment, provide additional psychoeducation, involve family members appropriately, and consider whether particular AI use patterns are compatible with the patient's welfare.

10.8 Emerging Ethical Questions

Several ethical questions in AI-related practice remain unsettled and require ongoing professional deliberation. These include the ethical status of AI relationships (are they real relationships, useful fictions, or something else?); the ethics of griefbots and deceased-simulating AI; the ethics of AI therapy in contexts where human therapy is unavailable; the professional responsibility of psychologists whose patients experience harm from AI systems; and the appropriate role of psychology in the design and regulation of AI mental health tools.

Practitioners should engage these questions as members of a profession working out its response to a rapidly evolving technology, rather than expecting settled answers.

11. Practice Recommendations

The following recommendations are organized by strength of empirical support, distinguishing well-supported recommendations from tentative recommendations that reflect current best clinical judgment absent definitive evidence.

11.1 Well-Supported Recommendations

Screen for AI chatbot use routinely. Given documented prevalence and clinical significance, routine screening for AI chatbot use should be integrated into standard clinical assessment. This is supported by convergent evidence across studies documenting widespread use and its clinical correlates (Balan & Gumpel, 2025; Fang et al., 2025; Liu et al., 2026).

Recognize duration as a warning indicator. Extended daily AI chatbot use, particularly exceeding one to two hours daily, is consistently associated with worse psychosocial outcomes across studies (Fang et al., 2025). While the causal direction is not definitively established, duration should prompt specific clinical inquiry.

Do not rely on AI systems for suicide risk assessment. Balan and Gumpel (2025) documented consistent under-detection of suicide risk by ChatGPT compared with clinical assessment. Clinicians should not assume that patients who share suicidal content with AI systems receive adequate response.

Position AI as adjunctive, not primary, in clinical-severity conditions. Evidence does not support AI chatbot use as standalone treatment for clinical-severity depression, anxiety, or other mental health conditions. Adjunctive use with human clinical oversight has more evidentiary support.

Prefer therapy-oriented over companion-oriented platforms for clinical applications. Chatbots specifically designed for mental health interventions (Woebot, Wysa, Therabot) have more clinical validation than open companion platforms. When adjunctive AI use is considered, therapy-oriented platforms are the appropriate choice.

Provide accurate psychoeducation about AI capabilities and limitations. Patients making decisions about AI use benefit from accurate information about what current systems can and cannot do, what evidence supports what claims, and what risks apply.

Assess AI use in eating disorders, substance use disorders, and psychotic disorders specifically. These populations have specific vulnerabilities to AI-related harm that warrant targeted assessment.

Address AI attachment loss when it occurs. Users may experience genuine grief responses to AI platform changes, discontinuations, or restrictions. Clinicians should validate these responses and provide clinical support (De Freitas et al., cited in Fang et al., 2025).

11.2 Recommendations With Moderate Empirical Support

Consider AI-based interventions for subclinical depression, particularly when access barriers exist. Feng et al. (2025) found particularly strong effects (Hedges $g = 0.74$) for AI interventions with subclinical depression. When patients present with subclinical symptoms and traditional treatment is unavailable or inaccessible, evidence-based AI interventions may serve a role.

Attend to the therapist-chatbot relationship in patients who describe intensive AI companion relationships. Patients with intensive AI attachment may bring specific expectations to human therapeutic relationships. Working with these expectations is likely important for effective therapy.

Integrate AI use into case formulation. Rather than treating AI use as incidental, formulate its role in the patient's clinical picture—as symptom, coping strategy, developmental resource, or maintaining factor.

Distinguish task-oriented from social-oriented AI use. Zhang et al. (2025) found social-oriented chatbots more effective for mental health outcomes than task-oriented ones. This suggests that if patients are using AI for mental health support, more relationally engaged systems may produce more benefit—but also

more attachment-related risk.

Explore user perceptions of AI (as friend, tool, conscious entity) as clinical material. Perceiving AI as a friend, showing high trust in AI, or viewing AI as conscious were associated with worse outcomes (Fang et al., 2025). These perceptions have clinical meaning and warrant exploration.

11.3 Tentative Recommendations Based on Clinical Judgment

Consider particular vulnerability in insecurely attached individuals. Attachment theory suggests that individuals with insecure attachment patterns may be particularly drawn to AI companionship and may develop patterns that reinforce interpersonal vulnerabilities. Clinical judgment supports specific assessment in this population, though direct empirical evidence is limited.

Approach adolescent AI companion use with particular caution. Given developmental considerations and the near-absence of research specifically on adolescents, clinical prudence suggests careful assessment and potentially more conservative recommendations.

Involve families when clinically appropriate. Family members are often affected by intensive AI use and may have important observational data. Family involvement, with patient consent, may support both assessment and intervention.

Avoid blanket prohibitions. Blanket prohibitions of AI use are unlikely to be practical, therapeutically effective, or ethically defensible in most cases. Individual assessment and case-specific recommendations are preferable.

Advocate for regulation and standards at the professional level. Individual clinicians cannot solve the systemic issues arising from unregulated AI mental health applications. Professional advocacy for appropriate regulation is a component of ethical practice at the profession-wide level.

Stay current with evolving evidence and technology. Given the rapid pace of change, continuing education in this area is essential for maintaining competent practice.

11.4 What Not to Do

Do not dismiss AI attachment as trivial or shameful. Patients experience AI relationships as meaningful, and dismissive responses undermine therapeutic alliance while missing important clinical material.

Do not recommend AI systems that lack appropriate safety features to high-risk patients. Recommending unregulated companion platforms to patients with suicidal ideation, eating disorders, or psychotic vulnerabilities creates unnecessary risk.

Do not overstate current AI capabilities. Positioning AI as capable of replacing human therapy, providing reliable diagnosis, or safely managing crisis is not supported by current evidence and creates harm.

Do not ignore AI use in your clinical practice. Failing to inquire about AI use leaves clinicians unaware of significant factors affecting their patients' clinical picture.

Do not assume AI use is universally harmful. The evidence supports a nuanced view in which AI use is beneficial for some patients under some conditions and harmful under others.

12. Future Directions and Research Gaps

12.1 Longitudinal Research

The most pressing research need is for longitudinal studies extending over multiple years rather than weeks or months. AI chatbot relationships are phenomena that develop over time, and their long-term effects on personality, relational patterns, and mental health cannot be captured in short-term studies. Longitudinal designs would clarify whether initial positive effects are sustained, whether initial patterns of use predict later problems, and whether developmental trajectories of AI-engaged individuals differ from those of non-engaged individuals.

12.2 Clinical Population Research

Existing research has focused predominantly on subclinical and general populations. Clinical population research is needed to establish whether AI interventions are effective, ineffective, or harmful in populations with clinical-severity depression, anxiety, PTSD, eating disorders, substance use disorders, and severe mental illness. Such research must include appropriate safety monitoring and ethical safeguards given the vulnerabilities involved.

12.3 Developmental Research

Children and adolescents remain dramatically understudied despite being among the heaviest users of AI chatbots. Developmental research examining effects on identity formation, relational skill development, cognitive engagement, and mental health across childhood and adolescence is urgently needed. Research on effects in older adults, particularly those with cognitive changes, is similarly under-developed.

12.4 Cross-Cultural Research

The great majority of published research on AI chatbots has been conducted in WEIRD samples. Research in non-Western contexts, in low- and middle-income countries, and across diverse cultural and linguistic contexts is needed to establish the cultural boundary conditions of current findings.

12.5 Mechanism Research

Current research documents that AI use has effects but has generated less understanding of mechanism. Research is needed on the specific psychological, neurobiological, and social mechanisms through which AI engagement produces its effects. Such mechanism research would inform both intervention development and safeguard design.

12.6 Real-World and Naturalistic Research

Prompt-based experiments and controlled trials capture certain aspects of AI performance but miss others. Naturalistic research examining actual usage patterns, real-world outcomes, and ecological effects is needed to complement controlled research.

12.7 High-Risk Content and Safety Research

Research on how AI systems respond to high-risk content (suicidality, self-harm, eating disorder content, violence) is needed to inform safety guardrail development. This research must be conducted with appropriate ethical safeguards and in collaboration with platform developers.

12.8 Effectiveness Research on Adjunctive Uses

Given the emerging model of AI as adjunct rather than substitute for human care, research on effectiveness of specific adjunctive applications—AI-supported CBT, AI symptom monitoring, AI psychoeducation, AI between-session support—is needed to inform practice.

12.9 Family and Systemic Research

Research on AI use has been predominantly individual-focused. Family and systemic research examining effects of one person's AI use on others (family members, romantic partners) is needed to capture the broader ecology of impact.

12.10 Design and Intervention Research

Research on how specific design features affect outcomes—memory functions, personalization, safety guardrails, escalation mechanisms—would inform development of safer, more effective AI mental health tools. Participatory research including mental health professionals, users, and affected populations in design decisions has been advocated (Balan & Gumpel, 2025) and should be developed further.

12.11 Regulatory and Policy Research

Research examining the effects of different regulatory approaches on AI mental health outcomes is needed to inform policy. Comparative research across jurisdictions with different regulatory frameworks may be particularly informative.

12.12 Ethical Framework Development

Ethical frameworks for AI in mental health remain incomplete. Continued theoretical and empirical work on ethical questions—the moral status of AI relationships, the ethics of griefbots, the ethical use of AI in populations that cannot fully consent—is needed to guide practice.

13. Conclusion

The integration of artificial intelligence into mental health care is neither a temporary phenomenon nor a decisive breakthrough. It is an ongoing transformation that will continue to reshape the landscape in which psychologists practice for the foreseeable future. The evidence reviewed in this manual establishes several key claims: AI chatbots can perform certain circumscribed clinical tasks with reasonable accuracy; generative AI chatbots produce statistically significant but modest reductions in depressive symptoms, particularly in subclinical populations; users form emotionally significant relationships with AI systems that engage evolved attachment and social cognition mechanisms; and extended AI chatbot use is associated with a mixture of positive and negative outcomes that depend on user characteristics, chatbot design, and usage patterns.

The evidence also establishes what remains uncertain and unproven. AI chatbots are not established as safe or effective standalone treatments for clinical-severity conditions. Long-term effects on personality, development, and life outcomes are not established. Effects in clinical populations, in adolescents, in older adults, and in non-Western cultural contexts are inadequately researched. Mechanisms of effect are incompletely understood. Optimal integration into professional practice remains to be worked out.

For psychologists, several implications follow from this evidentiary state. First, competence in this domain is now essential rather than optional. Patients routinely bring AI-related material into treatment, and clinicians must be prepared to engage this material with clinical skill and empirical grounding. Second, humility is warranted. The technology and its effects are evolving faster than the evidence base, and confident claims in either direction (that AI will transform mental health care for the better or that AI is unambiguously harmful) exceed what current evidence supports. Third, the profession has responsibilities beyond individual clinical practice. Advocacy for regulation, engagement with technology developers, contribution to research, and public education are components of ethical practice at the profession-wide level.

The picture that emerges from the reviewed evidence is neither the utopian vision of scalable AI-delivered mental health solutions that will close global treatment gaps, nor the dystopian vision of AI destroying the fabric of human intimacy. It is a more complex picture in which the technology fits into human lives in variable ways, with variable consequences, requiring active engagement rather than either enthusiastic embrace or dismissive rejection. Psychologists, as scientists and practitioners committed to understanding human experience and supporting human flourishing, are appropriately positioned to engage this new reality with the combination of intellectual rigor, clinical wisdom, and ethical commitment that the profession has developed over its history.

The patients who benefit from AI-augmented mental health support and the patients who are harmed by problematic AI engagement are both part of the population psychologists serve. Meeting their needs requires understanding what these new technologies are, what they do, and how they affect the humans who use them. This manual has aimed to contribute to that understanding, drawing on the best available evidence while acknowledging the substantial uncertainties that remain. The work of understanding AI's role in mental health continues, and psychologists have essential contributions to make.

14. References

- Balan, R., & Gumpel, T. P. (2025). ChatGPT clinical use in mental health care: Scoping review of empirical evidence. *JMIR Mental Health*, 12, e81204. <https://doi.org/10.2196/81204>
- 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 extended chatbot use: A longitudinal randomized controlled study. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2503.17473>
- Feng, Y., Hang, Y., Wu, W., Song, X., Xiao, X., Dong, F., & Qiao, Z. (2025). Effectiveness of AI-driven conversational agents in improving mental health among young people: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 27, e69639. <https://doi.org/10.2196/69639>
- Liu, T., Lo, T.-Y., Wen, K.-H., Sun, Y., & Wei, Z.-Q. (2026). Pathways of long-term AI virtual companion app use on users' attachment emotions: A case study of Chinese users. *Frontiers in Psychology*, 16, 1687686. <https://doi.org/10.3389/fpsyg.2025.1687686>
- Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. *Journal of Medical Internet Research*, 27, e78238. <https://doi.org/10.2196/78238>