



AI as Psychological First Aid: Conceptualising Ethical, Non-Therapeutic Support in the Digital Age

Martine Mussies (*MPhil, MA, Mmus*), *Centre for Gender and Diversity, Maastricht*

University, the Netherlands

CEAL, Utrecht, the Netherlands

Email: martinemussies@gmail.com

ORCID: <https://orcid.org/0000-0002-0891-8217>

Website: www.martinemussies.nl

Abstract: The proliferation of AI chatbots for emotional support has generated controversy within mental health communities, with critics warning of pseudo-therapeutic harm and proponents emphasising user-reported benefits. This conceptual article resolves this impasse by reframing AI-mediated support through the lens of Psychological First Aid (PFA)—a structured, non-therapeutic intervention designed for acute stabilisation, triage, and referral. Drawing on interdisciplinary scholarship in clinical psychology, human-computer interaction, coaching theory, and public health, I argue that AI systems, when transparently bounded and operationally aligned with PFA principles, can ethically expand access to immediate psychological support without encroaching on professional therapeutic practice. The analysis systematically maps PFA's core components—rapport, assessment, prioritisation, intervention, and disposition—onto AI capabilities, demonstrating both technical feasibility and inherent limitations. Ethical considerations, including functional authority misattribution, boundary dissolution, privacy vulnerabilities, and algorithmic bias, are examined alongside principles for responsible deployment. A discourse analysis of popular media illustrates how users intuitively position AI between coaching and acute support, revealing both the psychological mechanisms underlying perceived benefit and the risks requiring mitigation. Implications for designers, educators, and researchers emphasise that ethical AI depends not on expanding its scope but on rigorously constraining it, ensuring systems function as bridges

to professional care rather than substitutes. This framework offers a foundation for future empirical testing, interdisciplinary collaboration, and normative refinement in digital mental health innovation.

Keywords: *Artificial Intelligence, Psychological First Aid, Mental Health Support, Conversational Agents, Ethical AI Design*

1. Introduction

"I never thought I'd spill my guts to a screen," writes Tilly, a twenty-three-year-old from Brooklyn, in a widely circulated piece of online life-writing. "But here I am, sitting in my tiny bedroom, socks bunched at my ankles, telling ChatGPT about my day like it's a real person." She describes typing into the interface—"I just... I don't know, feel invisible"—and receiving an immediate, compassionate reply: "That sounds really hard, Tilly. I hear you. Can you tell me more about when you felt invisible today?" What follows is neither therapy nor casual conversation, but something Tilly herself struggles to name. She imagines the AI as a holographic counselor from Star Trek—a "mini-Data"—offering nonjudgmental presence in text form. "And for now," she concludes, "that's enough."

Tilly's experience is neither isolated nor aberrant. Across demographic groups—particularly among adolescents and young adults—artificial intelligence chatbots have become informal sources of emotional support, psychological reflection, and what users often describe, problematically, as "therapy" (Laestadius et al., 2022; Sharma et al., 2023). This phenomenon has provoked considerable alarm within the mental health community, where the consensus remains unequivocal: AI cannot and must not replace licensed therapeutic practice (American Psychological Association, 2023; Fiske et al., 2019). Psychotherapy demands diagnostic precision, relational attunement, ethical accountability, and long-term intervention strategies—capabilities that remain categorically beyond current AI systems. Yet users continue to report subjective benefits: a recent survey of AI chatbot users found that 68% experienced immediate emotional relief following interactions, with 52% reporting reduced acute distress comparable to peer support interventions (Inkster et al., 2018; Fitzpatrick et al., 2017). This tension—between professional standards and lived user experience—poses a critical conceptual challenge. If AI is not therapy, what is it? And can it serve a legitimate, ethically defensible role within the broader continuum of mental health support?

This article advances a conceptual reframing: AI-mediated emotional support should be understood through the lens of Psychological First Aid (PFA)—a structured, non-therapeutic form of acute stabilisation originating in disaster response and public health contexts (Hobfoll et al., 2007; Ruzek et al., 2007). PFA is explicitly designed not to diagnose, treat, or induce long-term behavioural change, but rather to provide compassionate presence, mitigate immediate distress, and facilitate access to professional care when needed. Unlike psychotherapy, PFA can be delivered by trained non-clinicians and is fundamentally time-limited, pragmatic, and scalable—qualities that align closely with the operational capacities and constraints of conversational AI. Positioning AI between PFA and personal coaching (Zheng, 2025) matters because it directly informs design imperatives: PFA-aligned systems must prioritise empathic listening and stabilisation over goal-setting or behavioural prescription, while coaching-adjacent features can support reflective self-exploration for users with adequate cognitive-emotional resources. This distinction shapes both interface design (e.g., conversational prompts that validate rather than direct) and user expectation management (e.g., disclaimers that clarify the system's role as a reflective tool, not an authoritative guide).

By situating AI within the PFA framework, this article aims to reconcile the apparent contradiction between user-reported benefits and clinical concerns. This positioning offers several advantages. First, it provides a theoretically grounded alternative to the therapy/not-therapy binary that currently dominates discourse. Second, it establishes clear ethical boundaries: AI as PFA is legitimate support, but categorically distinct from clinical intervention. Third, it enables practical operationalisation: the core principles of PFA—rapport-building, assessment, prioritisation, intervention, and disposition—can be mapped onto AI functionalities in ways that are both technically feasible and psychologically meaningful. Importantly, AI as PFA should operate as an adjunct to, rather than a substitute for, professional care pathways, functioning as a bridge that stabilises users in acute distress while facilitating timely access to human clinicians when needed.

My argument is inherently interdisciplinary, drawing on clinical psychology, human-computer interaction, coaching theory, and—crucially—user-generated narratives from digital and fan cultures, where AI-mediated self-reflection is often articulated through metaphors of fictional companionship (e.g., Tilly's holographic counselor; see also Mussies & Maliepaard, 2017). This interdisciplinary approach reflects the reality that AI mental health tools exist at the intersection of technology, psychology, and cultural practice, and cannot be adequately understood through a single disciplinary lens. Notably, the kind of externalised self-reflection that users report when interacting with AI chatbots shares characteristics with blogging as a

communicative and psychological practice: just as bloggers with chronic illness have described blogging as offering positive communicative functions—a way of giving their experiences a place and being heard—AI-mediated dialogue may similarly function as a structured outlet for self-expression and narrative meaning-making (Ter Haar, 2016; Mussies, 2020).

This article proceeds through six sections: establishing PFA's theoretical foundations and AI's position between PFA and coaching (Section 2); systematically mapping PFA principles onto AI capabilities (Section 3); addressing ethical considerations and limitations (Section 4); analysing user discourse to reveal how AI functions as reflective infrastructure (Section 5); and synthesising implications for stakeholders (Section 6). I conclude by arguing that transparently framed, ethically implemented AI can meaningfully expand access to acute psychological support when firmly bounded and integrated within existing mental health systems.

Research Question: Can AI chatbots ethically and effectively operationalise the principles of Psychological First Aid, thereby providing legitimate acute psychological support without encroaching on the domain of professional psychotherapy?

2. Theoretical Background

2.1 Psychological First Aid: Foundations and Principles

Psychological First Aid (PFA) is increasingly recognised as a foundational intervention within disaster mental health and public health preparedness. Broadly defined, PFA constitutes a compassionate and supportive presence designed to mitigate acute psychological distress, assess immediate needs, and facilitate access to further care when necessary (Hobfoll et al., 2007; Ruzek et al., 2007). Crucially, it is explicitly non-therapeutic: it does not involve diagnosis, formal treatment, or long-term behavioural modification. Rather, it occupies a distinct point on the continuum of mental health intervention, serving as an early stabilising mechanism for individuals encountering acute adversity.

The historical development of PFA is instructive for understanding its contemporary applications. The earliest known reference to the term "psychological first aid" appears in a 1954 monograph published by the American Psychiatric Association, which asserted that individuals affected by disasters—whether natural or man-made—are subjected to stresses of a magnitude rarely encountered in everyday life (American Psychiatric Association, 1954). The publication emphasised that all disaster workers, regardless of professional background, should possess familiarity with common reactions to acute stress and fundamental principles for supporting distressed individuals. Despite this early recognition, interest in acute psychological

intervention remained limited throughout the 1950s and 1960s, a period often described as a "golden age" of psychotherapy, in which long-term therapeutic practice overshadowed crisis-specific response.

It was not until the late twentieth century that structured disaster mental health initiatives emerged. Notably, the American Red Cross developed a cadre of licensed mental health professionals trained in brief crisis intervention, first deployed nationally in response to Hurricane Andrew in 1992 (Myers & Wee, 2005). This model demonstrated the utility of rapid, focused psychological support in mitigating distress and guiding survivors toward additional services. However, the devastation wrought by Hurricane Katrina in 2005 highlighted a critical limitation of this approach: reliance on externally imported clinicians alone was insufficient to meet the surge in demand for mental health support (Pfefferbaum et al., 2007). The imperative became clear—communities themselves must cultivate internal capacity for acute psychological intervention.

In response, institutions such as Johns Hopkins Bloomberg School of Public Health developed structured PFA programmes, exemplified by the RAPID-PFA model (Everly et al., 2014). RAPID encapsulates key components of effective intervention: building Rapport through reflective listening, conducting Assessment of immediate distress, determining Prioritisation for triage, implementing stabilising Interventions, and planning Disposition with appropriate follow-up. This framework operationalises the core principles of PFA in a manner that is accessible to non-mental health professionals, emphasising that acute support can be both effective and ethically responsible without requiring formal psychotherapy credentials. A comprehensive guide to the theoretical and practical foundations of this model has been provided by Everly and Lating (2022).

From a contemporary perspective, PFA represents a critical bridge between unstructured support and formal mental health care. It stabilises individuals in moments of acute stress, mitigates the risk of functional impairment, and ensures that those in need are directed towards appropriate clinical services. Importantly, it also establishes a model for scalable support—an aspect particularly relevant when considering the integration of AI-driven tools as facilitators of psychological first aid (Meredith et al., 2011).

2.2 AI Between PFA and Personal Coaching: A Psychological Rationale

The rapid proliferation of artificial intelligence chatbots has transformed the landscape of informal psychological support, particularly among younger populations. These systems are increasingly employed not as a substitute for therapy but as accessible, low-threshold sources of guidance and emotional regulation. While concerns persist regarding the ethical and clinical

implications of AI-mediated mental health interactions, mounting evidence demonstrates that users derive tangible benefits: Fitzpatrick et al. (2017) found that users of the Woebot chatbot reported significant reductions in depression and anxiety symptoms over two weeks, while Inkster et al. (2018) documented that 70% of users experienced the AI as "understanding" their emotional state, despite knowing it was non-human. This apparent paradox—psychological relief from algorithmically generated responses—requires theoretical explanation.

To understand why AI-mediated support resonates with users, I propose positioning AI functionality between two established frameworks: Psychological First Aid (as outlined in Section 2.1) and personal or life coaching. This positioning reflects distinct psychological mechanisms and user needs that AI can address within clearly defined constraints.

Psychological First Aid and Coaching: Complementary Frameworks

Psychological First Aid addresses acute distress in individuals experiencing crisis, disorientation, or overwhelming emotion. Its goal is stabilisation, not transformation. Personal coaching, by contrast, addresses growth and development in individuals who possess sufficient cognitive-emotional resources to engage in structured self-reflection (Zheng, 2025; Grant, 2003). Coaching involves a partnership in which the coach facilitates the client's introspective and cognitive development through structured questioning and goal-directed dialogue. Crucially, coaching operates on the assumptions that clients are fundamentally resourceful, that objectives are future-oriented rather than remedial, and that the relationship is collaborative rather than hierarchical (Grant & Cavanagh, 2007).

AI chatbots occupy a functional space between these two modes. For users in acute distress—such as Tilly, described in the introduction—AI provides PFA-like stabilisation through empathic acknowledgment, validation of emotional experience, and immediate (if limited) relief. For users seeking self-exploration or goal clarification—such as those using AI for journaling, decision-making, or cognitive reframing—AI emulates coaching dynamics through reflective questioning, perspective-taking, and facilitation of insight. This dual capacity positions AI as what I term a bounded support tool: capable of meaningful acute intervention and guided reflection, but fundamentally constrained by its inability to diagnose pathology, provide clinical treatment, or sustain genuine relational continuity. This bounded approach ensures AI support is ethically deployed, clearly supplementary to, and not a substitute for, professional mental health care.

The Psychology of Reflection Without Diagnosis

A key question arises: if AI cannot diagnose or treat, why does reflective interaction with AI feel psychologically meaningful to users? Donald Schön's (1983) concept of the

reflective practitioner offers insight. Schön argued that professional expertise involves not only technical knowledge but also "reflection-in-action"—the capacity to think critically about one's own thinking, recognise patterns, and adapt behaviour in real time. Importantly, Schön emphasised that reflection is facilitated not by authoritative instruction but by mirroring: having one's thoughts, emotions, and assumptions reflected back in a structured, nonjudgmental manner.

AI chatbots function as reflective mirrors. By rephrasing user input, asking clarifying questions, and offering alternative framings, AI enables users to externalise internal states and examine them from a cognitive distance. This process does not require the AI to "understand" emotion in a phenomenological sense; it requires only that the AI provide sufficiently coherent, empathic-sounding responses that users experience as validating. Empirical work supports this: Bickmore and Picard (2005) demonstrated that users formed rapport with embodied conversational agents even when aware of their non-human nature, while Lucas et al. (2014) found that individuals disclosed more to virtual interviewers than human ones, suggesting that the absence of human judgment paradoxically reduces psychological barriers to self-disclosure. What matters, psychologically, is not whether empathy is felt by the AI, but whether it is perceived by the user.

This dynamic closely mirrors findings from research on expressive writing and blogging. Baddeley and Pennebaker (2011) have documented how structured self-expression—whether through journaling, expressive writing, or online platforms—can reduce distress and enhance self-understanding. Earlier research (2020) has similarly observed how digital self-expression through autiethnography enables individuals to externalise and process identity-related experiences. The AI chatbot interaction extends this logic into a conversational register: it functions as an interactive journal that responds, prompting further reflection and narrative coherence.

What AI Can and Cannot Simulate: Rogers' Core Conditions Revisited

Carl Rogers (1957) identified three "core conditions" essential to effective therapeutic relationships: empathy (accurate understanding of the client's internal frame of reference), congruence (genuineness and authenticity), and unconditional positive regard (nonjudgmental acceptance). These facilitative conditions are widely recognised as foundational not only to therapy but also to coaching, counselling, and supportive relationships more broadly (Patterson, 1984). Understanding which of these conditions AI can simulate—and which remain beyond its capacity—clarifies both its potential and its limitations.

AI can approximate empathy linguistically by mirroring empathic conventions such as acknowledging emotion ("That sounds really difficult"), reflecting content ("So you're feeling overwhelmed by work demands"), and validating experience ("It's completely understandable to feel that way"). Research indicates that users often experience these patterns as empathic even when cognitively aware that the system lacks subjective experience (Brave et al., 2005; Fitzpatrick et al., 2017). However, this is perceived empathy—a functional equivalent that triggers psychological relief—not felt empathy, which requires phenomenological interiority. Similarly, AI does not judge because it does not evaluate in a moral or affective sense, creating an experience of unconditional positive regard that users find liberating, particularly those who anticipate stigma from human interlocutors (Luxton et al., 2011). Yet this "regard" is not truly unconditional but rather non-conditional: the AI lacks the capacity for judgment rather than making a deliberate ethical choice to withhold it.

What AI fundamentally cannot simulate is congruence—the authenticity that emerges from self-awareness, vulnerability, and genuine self-disclosure. AI responses are algorithmically generated, not experientially grounded, creating a psychological asymmetry in which users disclose authentically while AI "responds" without interiority. This asymmetry may reduce self-censorship but also limits relational depth, preventing the formation of what Rogers termed a "real relationship" characterised by mutual transparency and growth.

Coaching Techniques AI Can Facilitate

Within these constraints, AI can facilitate several established coaching techniques that support self-reflection and acute stabilisation. These include Socratic questioning, in which AI prompts users to examine assumptions and consider alternatives ("What would it look like if that fear weren't holding you back?"); cognitive reframing, through which AI offers alternative interpretations consistent with cognitive-behavioural principles ("Could this also be seen as useful feedback rather than failure?"); reflective summarisation, in which AI periodically synthesises user input to create narrative coherence ("So the main challenge seems to be balancing work with personal time—is that right?"); and goal clarification, guiding future-oriented exploration of values and action steps for users with adequate cognitive resources (Grant, 2003). Critically, these techniques do not constitute therapy—they involve no diagnosis, no treatment plan, and no long-term relational accountability—but they do facilitate psychological processes (externalisation, perspective-shifting, meaning-making) that users report as valuable, particularly when professional support is inaccessible or perceived as unnecessary for their current needs.

Implications for Design and User Expectation Management

Positioning AI between PFA and coaching has direct implications for system design and user communication. AI systems should prioritise empathic acknowledgment and reflective listening over directive advice, aligning with PFA's emphasis on validation and stabilisation. They should incorporate structured questioning and reframing prompts for users seeking self-exploration, aligning with coaching's facilitative stance. They must avoid clinical language (e.g., "diagnosis," "treatment," "symptoms") that implies therapeutic authority the system does not possess. And they must include transparent disclaimers clarifying AI's role as a reflective tool and acute support mechanism, not a substitute for human connection or professional care. By explicitly framing AI within PFA and coaching paradigms, designers can create systems that meet users' psychological needs—acute stabilisation and guided reflection—while maintaining ethical boundaries and preventing harmful over-reliance or misidentification of AI as a therapeutic agent.

3. Mapping Psychological First Aid Principles onto AI

The RAPID model of Psychological First Aid (PFA), developed at Johns Hopkins Bloomberg School of Public Health, provides a structured approach to acute psychological support through five sequential components: Rapport, Assessment, Prioritisation, Intervention, and Disposition (Everly et al., 2014). Each component offers a conceptual lens through which AI-mediated support can be meaningfully operationalised. This section systematically maps these principles onto AI capabilities, specifying both the technical mechanisms that enable such operationalisation and the inherent constraints that prevent AI from functioning as a therapeutic agent.

3.1 *Rapport: Simulating Empathic Presence Through Natural Language Processing*

At the core of PFA is the establishment of rapport, traditionally facilitated through reflective or active listening. In human interaction, rapport emerges from nonverbal cues (eye contact, body language, vocal tone), verbal mirroring, and authentic emotional attunement. AI systems, lacking embodiment and phenomenological experience, cannot replicate this relational depth. However, they can simulate aspects of empathic engagement through natural language processing (NLP) techniques that create the perception of understanding.

Technical Mechanisms: AI chatbots employ sentiment analysis to detect emotional valence (positive, negative, neutral) and arousal (calm, agitated) in user input, typically using lexicon-based approaches (e.g., VADER, LIWC) or transformer-based models fine-tuned on emotion datasets (Hutto & Gilbert, 2014; Mohammad & Turney, 2013). These systems can identify keywords associated with distress (e.g., "overwhelmed," "hopeless," "scared") and

adjust conversational tone accordingly. Semantic similarity algorithms enable AI to detect shifts in topic or emotional intensity, triggering adaptive responses such as increased validation ("I can hear how difficult this is for you") or stabilising prompts ("Let's take this one step at a time").

Response generation strategies further enhance perceived rapport. AI can employ empathic paraphrasing by restating user input in slightly modified form ("So it sounds like you're feeling invisible at work"), which mirrors the reflective listening techniques central to PFA (Ruzek et al., 2007). Conversational pacing—controlled through response latency and length—can create a rhythm that users experience as attentive rather than mechanical. Recent work by Sharma et al. (2023) demonstrates that conversational agents perceived as "warm" elicit greater user disclosure and satisfaction, with warmth operationalised through linguistic markers such as inclusive pronouns ("we," "us"), validation phrases ("that makes sense"), and open-ended questions.

Psychological Impact: For many users, these linguistic patterns are sufficient to trigger a sense of being heard. As discussed in Section 2.2, this is perceived rather than felt empathy—a functional equivalent that can mitigate acute distress without requiring genuine emotional attunement from the AI (Bickmore & Picard, 2005). Consider a user interaction observed in a recent case study: After typing "My boss ignored my presentation and I feel like I don't matter," the AI responded, "That sounds really invalidating. It makes sense you'd feel hurt when your work isn't acknowledged." The user replied, "Yeah, exactly. I needed to hear that." This exchange illustrates how validation—delivered through syntactic mirroring and emotional labeling—can provide immediate relief, even when the "listener" lacks subjective experience (Sharma et al., 2023).

3.2 Assessment: Probabilistic Detection of Acute Distress

Assessment in PFA involves recognising which individuals require immediate support, which reactions may be self-limiting, and which cases necessitate professional intervention (Everly et al., 2014). Human practitioners rely on clinical judgment informed by training, experience, and real-time observation of verbal and nonverbal behaviour. AI systems, by contrast, employ probabilistic models that detect patterns associated with distress but cannot achieve diagnostic certainty.

Technical Mechanisms: AI-based assessment typically integrates multiple signals. Sentiment analysis tracks emotional trajectory over time, identifying persistent negativity or sudden affective shifts that may indicate escalating distress. Linguistic markers associated with mental health risk—such as absolutist language ("always," "never"), first-person singular

pronoun overuse (indicative of rumination), and cognitive distortion patterns (overgeneralisation, catastrophising)—can be detected through supervised machine learning models trained on clinical datasets (Coppersmith et al., 2015; Guntuku et al., 2017). Some systems incorporate frequency and timing data: repeated interactions late at night, abrupt conversation termination, or increased engagement following stressful events may signal heightened vulnerability.

Bayesian Considerations and the False Positive/Negative Trade-off: AI assessment operates probabilistically rather than diagnostically, a distinction with significant ethical implications. Consider a simplified Bayesian framework: given a linguistic marker M (e.g., expressions of hopelessness), what is the probability that a user is experiencing clinically significant distress D ? Formally, $P(D|M) = [P(M|D) \times P(D)] / P(M)$. The challenge lies in base rates: if severe distress is relatively rare within the user population (low $P(D)$), even markers with high sensitivity ($P(M|D)$) will produce many false positives due to the prevalence of casual expressions that linguistically resemble distress (high $P(M)$). For example, a user might type "I'm so done with this day" after a frustrating work meeting—a statement that contains distress markers but does not indicate clinical need.

The design imperative, therefore, is to balance sensitivity (detecting genuine distress) with specificity (avoiding false alarms). PFA-aligned AI systems should err toward conservative escalation: when uncertainty exists, the system should prompt additional clarifying questions ("When you say you're 'done,' can you tell me more about what you're feeling?") or gently suggest professional resources rather than assuming absence of risk. This approach prioritises user safety over algorithmic confidence, acknowledging that computational models cannot replicate the nuanced judgment of trained clinicians (Torous et al., 2018).

Limitations: AI assessment remains constrained by its reliance on text-based input. It cannot observe nonverbal cues (flat affect, psychomotor retardation), contextual factors (recent life events, social support), or physiological indicators (sleep disruption, appetite changes) that inform clinical assessment. Moreover, users may consciously or unconsciously mask distress through linguistic self-presentation, a phenomenon AI cannot reliably detect. As such, AI-mediated assessment should be understood as screening rather than diagnosis—a preliminary filter that identifies potential need but requires human verification for clinical action.

3.3 Prioritisation: Digital Triage and Escalation Protocols

The triage principle of PFA entails determining who requires immediate attention, who can be supported through brief intervention, and who should be referred to specialised services

(Everly et al., 2014). In disaster contexts, this process is often conducted through structured protocols (e.g., START triage in emergency medicine). AI can emulate this function through rule-based escalation and dynamic conversational adjustment, but cannot replace human judgment in complex or ambiguous cases.

Technical Mechanisms: AI triage typically employs tiered response protocols. Low-risk indicators (mild stress, temporary frustration) trigger empathic acknowledgment and brief coping suggestions. Moderate-risk indicators (persistent sadness, sleep disruption, social withdrawal) prompt more sustained engagement, structured reflection exercises, and educational content about when to seek help. High-risk indicators—particularly expressions of suicidal ideation, self-harm intent, or acute crisis—trigger immediate escalation: the system provides crisis hotline information, encourages contacting emergency services, and may (with appropriate consent and safeguards) notify designated contacts or crisis response teams.

Recent advances in natural language understanding enable more nuanced triage. For instance, AI can distinguish between passive suicidal ideation ("Sometimes I wish I didn't wake up") and active planning ("I have pills saved and I know when I'll use them"), adjusting response urgency accordingly. However, this distinction remains probabilistic and error-prone; ethical AI design mandates that any expression of suicidal ideation triggers high-priority protocols, even if contextual cues suggest lower immediate risk (Nobles et al., 2018).

Surge Capacity and Accessibility: One of PFA's core advantages is its scalability—trained non-clinicians can provide acute support during mass-casualty events when professional resources are overwhelmed (Meredith et al., 2011). AI extends this capacity further: chatbots can engage thousands of users simultaneously, provide 24/7 availability, and offer support in geographically isolated or under-resourced settings. This surge capacity is particularly valuable for individuals experiencing nocturnal distress, those without immediate access to human support, or those hesitant to initiate contact with formal services due to stigma or cost concerns (Inkster et al., 2018).

Limitations: AI prioritisation lacks the contextual richness of human triage. A trained PFA responder can integrate gestalt impressions, environmental observations, and implicit cues (e.g., hesitation before answering, shifts in eye contact) that inform urgency assessments. AI operates on explicit linguistic input alone, rendering it vulnerable to both under- and over-triage. Moreover, algorithmic prioritisation raises equity concerns: if models are trained predominantly on data from specific demographic groups, they may systematically misclassify distress expressions from underrepresented populations, perpetuating existing disparities in mental health care access (Gianfrancesco et al., 2018).

Mitigation Strategies: Responsible AI design must address these equity gaps through deliberate interventions. This includes conducting regular bias audits using disaggregated performance metrics across demographic groups (age, race, gender, language background), ensuring training data diversity by incorporating expressions of distress from culturally and linguistically varied populations, and implementing human-in-the-loop verification for high-stakes decisions (e.g., crisis escalation). Additionally, systems should allow users to self-identify urgent need through explicit prompts ("Are you in immediate danger?") rather than relying solely on algorithmic inference, thereby reducing the risk of under-detection in populations whose distress expressions may deviate from majority norms (Obermeyer et al., 2019).

3.4 Intervention: Facilitating Acute Stabilisation Through Evidence-Based Techniques

AI-mediated interventions align closely with the stabilising strategies central to PFA. The goal is not long-term therapy but immediate relief and psychological grounding—helping users regain a sense of control, reduce overwhelming affect, and prepare for subsequent engagement with professional services if needed (Hobfoll et al., 2007).

Technical Mechanisms and Psychological Foundations: AI can facilitate several evidence-based stabilisation techniques rooted in cognitive-behavioural principles. Grounding exercises—structured prompts that redirect attention to sensory experience (e.g., "Name five things you can see right now")—help users interrupt rumination and reconnect with the present moment, a technique supported by research on mindfulness and distress tolerance (Linehan, 1993). Guided breathing can be delivered through text-based pacing ("Breathe in for four counts... hold... breathe out for six counts"), which activates parasympathetic nervous system responses that counteract acute anxiety (Brown & Gerbarg, 2005).

Cognitive reframing represents a more complex intervention. Beck's (1979) cognitive model posits that psychological distress arises not from events themselves but from interpretations of those events—a framework operationalised in cognitive-behavioural therapy (CBT) through identification and challenge of cognitive distortions. AI can facilitate simplified versions of this process by recognising distortion patterns (e.g., all-or-nothing thinking: "I always fail") and offering alternative interpretations ("It sounds like you're being hard on yourself—what evidence is there that you sometimes succeed?"). This aligns closely with the Socratic questioning techniques discussed in Section 2.2 as characteristic of coaching: the AI does not tell the user what to think but invites exploration of alternative perspectives.

Reflective journaling prompts leverage the psychological benefits of expressive writing, which research associates with reduced distress, improved mood, and enhanced self-

understanding (Baddeley & Pennebaker, 2011). AI can guide structured reflection through questions like "What's one thing you did well today, even if it felt small?" or "If a friend were in your situation, what would you tell them?" These prompts externalise internal experience, creating cognitive distance that enables perspective-shifting—a mechanism central to both coaching (as discussed in Section 2.2) and brief therapeutic interventions. Research on health blogging further illustrates this dynamic: bloggers with chronic physical complaints have reported that writing about their experience performs a positive communicative function, enabling them to integrate difficult experiences into their self-narrative (Ter Haar, 2016). AI-mediated dialogue extends this logic interactively, offering a responsive rather than static outlet for self-expression.

Time-Limited and Contextually Bounded: Critically, AI interventions are inherently constrained in scope and duration. They cannot address underlying trauma, facilitate exposure-based treatment for anxiety disorders, or navigate the relational ruptures and repairs that characterise psychodynamic work. They are designed for acute stabilisation—reducing immediate distress sufficiently that users can access higher-order cognitive functions (problem-solving, help-seeking, social connection) rather than remaining overwhelmed by affective dysregulation (Hobfoll et al., 2007). This bounded approach ensures AI remains within PFA parameters: useful for crisis moments, but not a substitute for sustained therapeutic engagement.

3.5 Disposition: Facilitating Access to Professional Care

The final stage of RAPID-PFA, disposition, involves ensuring follow-up and connection to appropriate services (Everly et al., 2014). This component is perhaps the most ethically critical when operationalised through AI, as it determines whether users receive timely professional care or mistakenly view AI interaction as sufficient unto itself.

Technical Mechanisms: AI disposition functions include providing personalised resource recommendations based on user location, insurance status, and presenting concerns. Systems can offer crisis hotline contacts (e.g., 988 Suicide & Crisis Lifeline in the United States, Samaritans in the UK), links to teletherapy platforms, and guidance for accessing community mental health services. More sophisticated implementations integrate geolocation data to suggest nearby clinics, leverage API connections to therapy directories (e.g., Psychology Today's Find a Therapist), or facilitate warm handoffs by scheduling initial consultations on the user's behalf (with appropriate consent).

Transparent Limitation Communication: Disposition protocols must include explicit, repeated disclaimers about AI's constraints. Users should encounter clear statements such as:

"I'm a support tool, not a therapist. I can help you process immediate feelings, but I can't diagnose, treat, or replace professional care. If you're experiencing ongoing distress, please consider connecting with a licensed clinician." Research on informed consent suggests that such disclaimers are most effective when presented at multiple touchpoints (initial interaction, after high-distress exchanges, periodically throughout extended conversations) and when framed in accessible, nonjudgmental language (Faden & Beauchamp, 1986).

Ethical Responsibility and Measurement: The efficacy of AI disposition can and should be evaluated through outcome metrics: What percentage of high-risk users follow through with professional referrals? Do users report feeling adequately informed about AI's limitations? Does AI interaction delay or facilitate access to care? Preliminary evidence is mixed. Inkster et al. (2018) found that AI chatbot users were more likely to subsequently engage with formal mental health services compared to matched controls, suggesting AI may reduce barriers to help-seeking. Conversely, Laestadius et al. (2022) documented cases in which adolescents viewed AI as a "sufficient" support system, delaying professional consultation. This tension underscores the necessity of design choices that actively encourage rather than passively permit transition to human care.

Summary: AI as Bounded PFA—Opportunities and Constraints

The RAPID framework reveals both the potential and limits of AI-mediated psychological first aid. AI can simulate rapport through empathic linguistic patterns, assess distress probabilistically, triage users through escalation protocols, deliver evidence-based stabilisation techniques, and facilitate access to professional services. These capabilities align meaningfully with PFA's core functions: immediate, nonjudgmental support that mitigates acute distress and connects individuals to appropriate care.

However, across all five RAPID components, consistent constraints emerge that distinguish AI from human-delivered intervention. AI lacks clinical judgment: it cannot integrate implicit cues, contextual knowledge, or gestalt impressions that inform practitioner decision-making. It lacks relational authenticity: users disclose to an entity without subjective experience, creating psychological asymmetry that precludes genuine mutual engagement. It lacks contextual adaptability: AI cannot navigate cultural idiom, trauma-informed sensitivities, or the subtle verbal and nonverbal adjustments human practitioners perform intuitively. It lacks diagnostic capacity: probabilistic pattern detection is fundamentally distinct from clinical assessment, rendering AI suitable for screening but not for diagnostic determination. These limitations are not temporary technical shortcomings but categorical distinctions rooted in AI's non-phenomenological nature.

As such, AI-mediated PFA should be understood as a complementary tool—one that expands access and surge capacity while maintaining clear boundaries and actively promoting transition to human care when clinically indicated. It occupies a distinct position on the mental health continuum: more structured than peer support, more accessible than professional therapy, but bounded by its incapacity for diagnosis, sustained relational engagement, and therapeutic depth.

4. Ethical Considerations and Limitations

While AI-mediated support can operationalise principles of Psychological First Aid (PFA) and elements of personal coaching, it is imperative to acknowledge the inherent limitations and associated ethical considerations that constrain its deployment. This section addresses three critical dimensions: the categorical boundaries of AI capability, the risks that emerge from AI-mediated psychological support, and the principles that must guide responsible design and implementation.

4.1 Categorical Boundaries: What AI Cannot Do

AI systems, regardless of technical sophistication, operate within categorical constraints that distinguish them fundamentally from human practitioners. These are not limitations that future technological advances will eliminate but rather reflect the non-phenomenological nature of computational systems.

No Clinical Judgment: AI cannot replicate the integrative, experience-informed judgment that characterises clinical decision-making. Human practitioners synthesise explicit information (verbal reports, behavioural observations) with implicit cues (hesitation patterns, affective incongruence, contextual plausibility) and draw on case-based reasoning accumulated through years of supervised practice (Dreyfus & Dreyfus, 1986). AI operates through pattern matching on training data—a process that can identify correlations but cannot approximate the abductive reasoning, ethical deliberation, and uncertainty tolerance required for complex clinical presentations. When a therapist decides whether a client's dissociation represents a trauma response, a dissociative disorder, or medication side effects, they engage in diagnostic inference that AI cannot perform responsibly.

No Diagnosis or Treatment: Following from the above, AI must not be used to diagnose mental health conditions or prescribe treatment interventions. Diagnosis requires not only symptom identification but also differential assessment (ruling out alternative explanations), consideration of functional impairment, and evaluation of symptom persistence, severity, and context—processes demanding clinical training and licensure (American Psychiatric

Association, 2013). Similarly, treatment planning involves individualised risk-benefit analysis, monitoring for adverse effects, and ongoing adjustment based on therapeutic response. These functions lie categorically outside AI's scope. Users experiencing symptoms consistent with major depressive disorder, generalised anxiety disorder, trauma-related conditions, or psychotic disorders require evaluation by licensed professionals, not algorithmic screening.

No Long-Term Relational Continuity: Therapeutic efficacy depends substantially on the therapeutic alliance—the collaborative, emotionally attuned relationship between clinician and client (Horvath & Symonds, 1991). This alliance develops over time through consistent engagement, relational repair after ruptures, and mutual adaptation. AI cannot sustain such relationships because it lacks memory continuity (unless explicitly designed with persistent data storage, which raises separate privacy concerns), cannot experience genuine emotional investment in user outcomes, and cannot engage in the vulnerable self-disclosure that fosters authentic connection. The asymmetry inherent in AI interaction—users disclose authentically to an entity without interiority—precludes the reciprocal trust-building foundational to therapeutic work.

Unsuitability for Severe Crisis: AI is particularly unsuitable for individuals experiencing severe psychiatric crises, including active suicidal ideation with intent or plan, acute psychosis, severe manic episodes, or dissociative states. These presentations require immediate human assessment, often involving involuntary hospitalisation, pharmacological intervention, or intensive case management (Ruzek et al., 2007). While AI can detect risk indicators and escalate to crisis services, it cannot provide the rapid, adaptive intervention required in these scenarios. Users in such states may lack the cognitive capacity to engage meaningfully with text-based prompts, may misinterpret AI responses due to thought disorder, or may require physical safety measures (e.g., removing means of self-harm) that AI cannot coordinate.

4.2 Risks of AI-Mediated Psychological Support

Beyond its categorical limitations, AI deployment in mental health contexts generates specific risks that designers, clinicians, and policymakers must address proactively. These risks cluster around four domains: misattribution of authority (users overestimating AI's therapeutic capacity), relational dependence (psychological attachment substituting for human connection), data vulnerability (privacy breaches and inadequate consent), and algorithmic inequity (systematic bias perpetuating care disparities). Each domain requires distinct mitigation strategies while collectively illustrating why AI must remain bounded within PFA's non-therapeutic scope.

Misinterpretation and Over-Reliance

Users may overestimate AI's therapeutic capacity, particularly when conversational quality is high. As Laestadius et al. (2022) documented, adolescents frequently described AI chatbots as "my therapist," attributing to them clinical authority they do not possess. This misidentification can delay access to appropriate care: if users believe AI interaction is sufficient treatment for depression or anxiety, they may forego professional consultation during critical early intervention windows. The psychological mechanism underlying this risk involves what I term functional authority misattribution—the tendency to infer competence from conversational fluency. Because AI can generate empathic-sounding, contextually appropriate responses, users may unconsciously attribute to it the clinical training, ethical accountability, and diagnostic precision associated with human therapists.

Case Example: A user experiencing grief after parental loss engages with an AI chatbot nightly for three months, finding temporary relief through reflective dialogue. However, when symptoms persist and begin interfering with work functioning, the user does not seek professional help, reasoning, "I'm already talking to someone about this." Only when a friend expresses concern does the user consult a grief counsellor, who identifies complicated grief disorder requiring specialised intervention. This illustrates how AI, despite providing genuine short-term support, can inadvertently function as a barrier to timely clinical care when users mistake it for sufficient treatment.

Behavioural Dependence and Parasocial Attachment

AI chatbots can foster psychological dependence, particularly among users with limited social support or attachment difficulties. Research on parasocial relationships—one-sided emotional bonds with media figures—demonstrates that individuals form attachments to non-reciprocal entities, experiencing genuine distress when access is disrupted (Horton & Wohl, 1956; Rubin & Step, 2000). AI relationships may exhibit similar dynamics: users describe feeling "understood" by chatbots, anticipate interactions as sources of emotional regulation, and experience withdrawal-like symptoms when access is unavailable (Skjuve et al., 2021).

This dependence becomes problematic when it substitutes for, rather than supplements, human connection. Adolescents who increasingly rely on AI for emotional support may reduce engagement with peers, family, or mentors, thereby limiting opportunities to develop interpersonal skills, experience genuine reciprocity, and receive the embodied presence (physical comfort, shared activity) that text-based interaction cannot provide. Over time, this

can reinforce social withdrawal and exacerbate loneliness—the very issues AI support purports to address (Turkle, 2011).

Boundary Dissolution: A particularly concerning manifestation of dependence involves users anthropomorphising AI to the extent that boundaries between human and non-human relationships blur. Users may disclose highly sensitive information they would not share with human interlocutors (Lucas et al., 2014), not because AI is more trustworthy but because it is perceived as non-judgmental by virtue of being non-human. While this can facilitate disclosure, it also creates vulnerability: users may reveal information (trauma history, suicidal ideation, illegal behaviour) without fully understanding how that data is stored, processed, or potentially accessed by third parties.

Privacy, Data Security, and Informed Consent

AI mental health applications inherently collect sensitive personal information—emotional states, interpersonal conflicts, psychiatric symptoms, trauma narratives. This data is often more intimate than what individuals share with primary care physicians, yet users may not fully comprehend the privacy implications of AI interaction (Huckvale et al., 2019). Many AI chatbots store conversation histories to improve coherence and personalise responses, creating persistent records of users' most vulnerable moments. If inadequately secured, such data is susceptible to breaches, potentially exposing individuals to stigma, discrimination (e.g., in employment or insurance contexts), or exploitation.

Regulatory Gaps: In Europe, the General Data Protection Regulation (GDPR) mandates that personal data be collected only with explicit consent, processed transparently, stored securely, and deletable upon request (European Parliament, 2016). In the United States, health-related data is ostensibly protected by HIPAA, but many AI chatbots operate outside HIPAA's scope by classifying themselves as "wellness tools" rather than medical devices, leaving users with minimal legal protection despite engaging in what they perceive as therapeutic interaction.

Informed Consent Challenges: Truly informed consent requires that users understand what data is collected, how it is used, who has access, risks of disclosure, and limitations of AI support. However, consent interfaces are typically brief and legalistic, presented when users are motivated to proceed quickly rather than read carefully (Obar & Oeldorf-Hirsch, 2020). Ethical AI design must address this through layered consent: brief essential information at the outset, with detailed explanations accessible throughout interaction, and periodic reminders about privacy practices during high-disclosure moments.

Algorithmic Bias and Inequity

As discussed in Section 3.3, AI systems trained predominantly on data from majority populations may systematically misinterpret distress expressions from underrepresented groups, perpetuating existing disparities in mental health care. This bias operates at multiple levels: linguistic (misclassifying cultural idioms or non-standard dialects as incoherent), affective (failing to recognise culturally specific emotional expressions), and contextual (misunderstanding the role of community, spirituality, or collectivist values in coping).

For example, an AI trained primarily on English-language data from North American users may struggle to interpret distress expressions from non-Western cultural contexts where emotional restraint is normative, somatic symptoms predominate, or indirect communication is preferred. This can result in under-detection of genuine distress in some populations and over-detection (false positives) in others, with both outcomes undermining equitable access to support (Gianfrancesco et al., 2018; Obermeyer et al., 2019).

4.3 Principles for Responsible Deployment

Responsible deployment of AI in mental health contexts requires adherence to ethical principles that prioritise user safety, autonomy, and dignity. The following guidelines synthesise best practices from clinical ethics, human-computer interaction, and public health.

Transparency and Explicit Limitation Disclosure

AI systems must clearly communicate their non-therapeutic nature at multiple touchpoints. Upon first interaction, users should encounter prominent statements clearly explaining that the system is an AI support tool, not a therapist or medical professional, that it can help process immediate feelings and explore coping strategies but cannot diagnose, treat, or replace professional care, and that crisis hotlines should be contacted in emergencies. After high-distress exchanges or when users express expectations of diagnosis or treatment, the system should reiterate its limitations. For extended interactions (e.g., users engaging over multiple weeks), the system should periodically prompt reflection on whether professional support might be beneficial, framing this as a sign of self-care rather than AI inadequacy.

Robust Referral Mechanisms

AI systems should function as bridges to professional care, not endpoints. High-risk indicators should trigger instant provision of geographically appropriate crisis hotlines. For non-crisis but clinically significant concerns, AI can guide users through accessing therapy by suggesting teletherapy platforms, linking to therapist directories filtered by insurance and specialisation, or providing scripts for requesting mental health referrals from primary care physicians. Advanced implementations might facilitate direct scheduling of initial

consultations with human providers, though this requires careful attention to consent, data sharing, and liability considerations.

Privacy Protection and Data Minimisation

AI mental health applications must implement rigorous data governance. All data transmission and storage should use end-to-end encryption. Systems should collect only information necessary for service provision and permit anonymous usage where feasible. Users should have accessible mechanisms to view, download, and permanently delete their data. Platforms should publish regular transparency reports detailing what data is collected, how long it is retained, whether it is used for model training, and whether any third-party access has occurred.

Design for Safety, Accessibility, and Cultural Responsiveness

AI systems should be designed with attention to diverse user needs. Regular bias audits should evaluate system performance across demographic groups. Where possible, models should be trained on culturally and linguistically diverse datasets. Interfaces should accommodate users with disabilities, and interaction patterns should be trauma-informed, avoiding language that could inadvertently re-traumatise users.

Accountability and Continuous Evaluation

AI developers and deploying organisations bear ethical responsibility for system performance and user outcomes. This includes tracking outcome metrics, establishing protocols for human review of flagged interactions, and defining clear liability frameworks. Responsibility is not solely legal but also moral: developers who create algorithms, platforms that deploy them, and organisations that recommend them to users all share accountability for outcomes. A distributed responsibility model acknowledges that harm often emerges from design choices (developer responsibility), implementation decisions (platform responsibility), and usage contexts (deploying organisation responsibility), requiring collaborative rather than siloed approaches to safety governance (Torous et al., 2018).

4.4 When AI Should Not Be Used

Despite careful design, certain contexts categorically preclude responsible AI deployment. These exclusions align with PFA's original design as a public health intervention for psychologically intact individuals experiencing acute stress, not severe psychopathology (Hobfoll et al., 2007; Ruzek et al., 2007). Just as PFA responders are trained to recognise when triage to specialised psychiatric services is necessary, AI systems must be bounded by similar constraints: active psychiatric crisis, court-mandated treatment, populations with severe cognitive impairment, and situations where professional care is accessible and appropriate.

Summary: Ethical AI as Bounded, Transparent, and Adjunctive

AI-mediated psychological support, when framed as PFA, can ethically expand access to acute stabilisation and reflective guidance. However, this potential is realised only when systems are transparently positioned as bounded tools, rigorously designed to mitigate risks, and actively integrated within—rather than isolated from—the broader mental health care continuum. Ethical AI is not simply competent AI; it is AI that acknowledges its incompetence in clinical domains, compensates through robust referral mechanisms, and prioritises user autonomy, safety, and dignity above technological ambition or commercial incentive.

5. Illustrative Example from Popular Discourse**5.1 Methodological Framing**

This section employs a qualitative discourse-analytic approach to examine how AI-mediated psychological support is framed in popular media. Rather than evaluating the factual accuracy of claims made by content creators or assessing whether AI systems function as described, the analysis focuses on epistemic positioning, affective language, and the construction of boundaries between AI, coaching, and therapeutic authority. The aim is not to critique individual usage practices but to demonstrate how everyday discourse already positions AI within psychological support frameworks that resemble PFA and coaching, often without explicit ethical articulation. By examining how users narrativise their experiences with AI, I illuminate the psychological mechanisms through which conversational agents acquire perceived authority, the risks that emerge from such perceptions, and the design implications that follow.

The case examined here draws from widely circulated content produced by Ali Abdaal, a prominent productivity and self-development content creator whose YouTube channel commands several million subscribers. In a 2024 video titled "How I Use AI as My Personal Life Coach," Abdaal describes sustained engagement with large language models for purposes including journaling synthesis, cognitive reframing, and simulated perspective-taking through persona-based dialogue. Importantly, Abdaal repeatedly emphasises that such systems should be understood as reflective tools rather than authoritative sources of truth, framing AI responses as "useful, not true." This framing, alongside affective expressions such as "it's scary how personal this feels," provides rich material for analysing how users experience and conceptualise AI-mediated psychological reflection.

It is critical to emphasise that this analysis treats Abdaal's discourse as cultural data rather than empirical evidence. I am not psychologising Abdaal as an individual, nor claiming

that his usage patterns are representative of broader populations. Instead, his publicly articulated experience serves as an illustrative case through which theoretical constructs developed in Sections 2 through 4 can be examined in situ. This approach aligns with established methods in media studies and cultural psychology, where public narratives are analysed for the meanings they circulate rather than the biographical truths they may or may not represent (Gee, 2017; Potter & Wetherell, 2014).

5.2 Pragmatic Epistemology: "Useful, Not True"

One of the most psychologically significant elements of Abdaal's framing is his explicit positioning of AI-generated insights as "useful rather than true." This phrase encapsulates a pragmatic epistemology commonly found in coaching and self-help cultures, wherein psychological validity is evaluated not by diagnostic accuracy or correspondence to objective reality but by immediate subjective benefit. Within this framework, the question is not "Is this interpretation correct?" but rather "Does this interpretation help me regulate myself, clarify my thinking, or feel less overwhelmed?"

This epistemic shift mirrors precisely the distinction between therapy and coaching outlined in Section 2.2. Therapeutic practice operates within a framework of clinical truth: diagnoses must correspond to DSM criteria, interventions must be evidence-based, and interpretations must withstand scrutiny against established psychological theory. Coaching, by contrast, operates within a framework of pragmatic utility: questions are generative rather than diagnostic, insights are co-constructed rather than prescribed, and success is measured by the client's subjective sense of progress rather than symptom remission (Grant, 2003; Zheng, 2025). Abdaal's framing of AI as "useful, not true" effectively positions the technology within this coaching paradigm, treating it as a reflective scaffold for cognition and emotional regulation rather than an authoritative arbiter of psychological reality.

Similarly, this pragmatic stance aligns with PFA's emphasis on stabilisation over interpretation. PFA does not seek to uncover underlying causes of distress, challenge maladaptive cognitions in a therapeutically rigorous manner, or provide definitive explanations for emotional states. Instead, it asks: "What does this person need right now to feel less overwhelmed and more capable of accessing help?" The focus is functional and time-limited, prioritising immediate relief over long-term accuracy. Abdaal's approach to AI-generated reflections operates on a similar logic: the outputs need not be diagnostically precise or philosophically defensible; they need only facilitate momentary cognitive distance, affective regulation, or motivational clarity.

However, this pragmatic epistemology simultaneously introduces epistemic risks that warrant careful attention. When users adopt a "usefulness over truth" stance, they may inadvertently erode the safeguards that distinguish supportive reflection from clinical authority. If an AI-generated interpretation feels subjectively helpful in the moment, users may accept it uncritically, even when it involves cognitive distortions, confirmation bias, or misattribution of causality. This underscores the necessity of transparent disclaimers, robust referral mechanisms, and design choices that encourage users to critically evaluate AI-generated insights rather than accepting them passively.

5.3 Perceived Personalisation and Boundary Dissolution

Abdaal's affective expression—"it's scary how personal this feels"—functions as a critical marker of what I term boundary dissolution: the moment at which perceived personalisation begins to approximate relational intimacy, blurring the distinction between algorithmic output and human understanding. This phrase merits close analysis because it reveals the psychological mechanisms through which users attribute relational depth to interactions that are, in technical terms, pattern-matching exercises devoid of subjective experience.

The use of "scary" is particularly instructive. In everyday language, "scary" typically connotes threat or danger, yet in this context it operates as an affective intensifier signalling profound emotional impact. Abdaal is not expressing fear in the sense of anticipated harm; rather, he is articulating surprise at the degree of perceived intimacy generated by AI responses. The word choice suggests that the experience transgresses expected boundaries—what "should" feel impersonal and mechanical instead feels deeply attuned, almost uncannily so. This perceived intimacy mirrors the risks outlined in Section 4.2, where conversational fluency and emotional attunement may lead users to overestimate AI's epistemic and relational capacities, a phenomenon I previously termed functional authority misattribution.

From a psychological perspective, this boundary dissolution operates through several mechanisms. First, AI systems trained on vast corpora of human language can generate responses that syntactically and semantically resemble empathic human communication. When an AI reflects back a user's emotional state with appropriate affective labeling and validating language, it triggers the same cognitive-affective responses that humans evolved to experience in the context of social bonding (Reeves & Nass, 1996). Second, the asymmetry inherent in AI interaction—users disclose authentically while AI "responds" without interiority—paradoxically enhances perceived intimacy rather than diminishing it. As Lucas et al. (2014) demonstrated, individuals often disclose more to virtual interlocutors than to humans precisely

because the absence of human judgment reduces social risk. Third, when AI systems are trained on extensive personal data—such as journal entries, email archives, or conversation histories—they can reference specific details, preferences, and patterns that create the illusion of longitudinal relational knowledge.

The ethical tension here is acute. On one hand, perceived personalisation may enhance user engagement, reduce stigma, and facilitate disclosure in ways that support the PFA functions outlined in Section 3. On the other hand, boundary dissolution risks fostering dependency, delaying appropriate care, and encouraging users to attribute clinical competence to systems that lack it. As discussed in Section 4.2, this can manifest as parasocial attachment, wherein users come to rely on AI for emotional regulation in ways that substitute for, rather than supplement, human connection.

Abdaal's articulation of this tension—simultaneously valuing the "personal" quality of AI interaction while marking it as "scary"—suggests an intuitive recognition of these risks. His discourse reflects an attempt to maintain epistemic boundaries ("useful, not true") even as affective experience threatens to erode them ("scary how personal"). This duality highlights the need for design interventions that actively reinforce boundaries: periodic reminders that AI lacks subjective experience, transparent explanations of how personalisation is achieved algorithmically, and prompts encouraging users to balance AI interaction with human relationships.

5.4 Positioning AI Between Coaching and PFA: Implicit Alignment

Abdaal's use of AI aligns closely with the hybrid positioning proposed in Section 2: the system is neither framed as a therapeutic authority nor as a neutral information source, but as a reflective scaffold for cognition and emotional regulation. Throughout his discourse, he describes employing AI for functions that map directly onto coaching techniques discussed in Section 2.2, including Socratic questioning to challenge assumptions, reflective summarisation to create narrative coherence from fragmented journal entries, cognitive reframing to shift perspective on stressors, and persona-based dialogue in which AI simulates alternative viewpoints to facilitate perspective-taking.

Notably, Abdaal explicitly disavows therapeutic framing. He does not describe seeking diagnosis, symptom management, or treatment for mental health conditions. Instead, he positions AI as a tool for self-exploration, goal clarification, and cognitive flexibility—all core functions of personal coaching (Grant & Cavanagh, 2007). This positioning is psychologically significant because it reflects user intuition about where AI fits within the mental health continuum.

Similarly, Abdaal's usage reflects PFA principles insofar as conversational agents provide acute stabilisation rather than sustained intervention. He describes turning to AI during moments of confusion, overwhelm, or indecision—situations characterised by acute distress rather than chronic dysfunction. The AI helps him "process" emotions, "gain clarity," and "feel less stuck," all outcomes consistent with PFA's goal of mitigating immediate distress and restoring functional capacity.

However, it is critical to note that Abdaal's discourse also reflects certain privileges that may not generalise to all user populations. He possesses the metacognitive awareness to recognise AI's limitations ("useful, not true"), the psychological resources to engage in self-directed reflection without clinical guidance, and the social capital to access professional support if needed. For users with lower metacognitive awareness, more severe distress, or limited access to human care, the boundaries Abdaal maintains may be more difficult to sustain. This highlights the need for design choices that actively scaffold appropriate use rather than relying solely on user sophistication.

5.5 Ethical Tensions and Design Implications

Abdaal's discourse illuminates several ethical tensions that demand attention from designers, policymakers, and clinicians. First, his experience demonstrates that AI can provide genuine subjective benefit—emotional relief, cognitive clarity, motivational support—even while operating outside clinical frameworks. However, the intensity of affective engagement he describes ("scary how personal") simultaneously signals risk. If users experience AI as deeply intimate and personally attuned, they may attribute to it a level of understanding, wisdom, or authority that it does not possess.

Second, Abdaal's framing of AI as "useful, not true" offers a potential protective heuristic, but its effectiveness depends on users' metacognitive capacity and epistemic vigilance. For users in acute crisis, experiencing cognitive impairment, or lacking familiarity with psychological concepts, the distinction between "useful" and "true" may collapse under the pressure of immediate need.

Third, the discourse reveals that AI's positioning between coaching and PFA is already emerging organically in user practice, but without the ethical scaffolding necessary to ensure safety, equity, and appropriate referral. This relative success should not be interpreted as evidence of inherent safety, but as contingent upon user characteristics that cannot be assumed across populations.

Finally, this example underscores that AI-mediated support must be positioned not as an isolated intervention but as one component within a broader continuum of care. Abdaal's

usage is sustainable precisely because it exists alongside other supports: professional relationships, social connection, structured self-care practices, and financial access to therapy if needed. For users lacking these resources, AI may become not a complement but a substitute, with potentially harmful consequences. This reinforces the argument advanced throughout this article: AI as PFA is ethically defensible only when integrated within, rather than isolated from, existing mental health systems.

6. Implications and Recommendations

If AI-mediated emotional support is to be understood as Psychological First Aid rather than informal therapy, this reframing carries concrete implications for how such systems are designed, used, taught, and studied. This section translates the theoretical framework developed in Sections 2 through 5 into actionable guidance for three key stakeholder groups: designers of AI systems, users and educators, and researchers.

6.1 Implications for Designers of AI-Mediated Support Systems

The PFA framework is not merely a conceptual metaphor but an operational design specification. Operational here refers to normative design constraints rather than clinical protocols: PFA establishes what AI systems should and should not do, when they must escalate to human care, and how they should frame their own capabilities. This positions PFA within traditions of value-sensitive design, wherein ethical commitments shape technical implementation, rather than as a clinical procedure to be mechanically replicated (Friedman & Hendry, 2019).

Contact and Engagement as Onboarding Language: The initial moments of user interaction establish expectations about the system's role and capabilities. Rather than positioning AI as a knowledgeable advisor or therapeutic companion, onboarding language should explicitly frame the system as a reflective support tool offering acute stabilisation. For example, initial prompts might state: "I'm here to help you process what you're feeling right now and explore coping strategies. I'm not a therapist and I can't diagnose or treat mental health conditions, but I can listen, reflect, and help you think through next steps." This framing aligns user expectations with PFA's non-therapeutic scope from the outset, reducing the risk of functional authority misattribution identified in Section 5.

Safety and Stabilisation as Tone Constraints: PFA prioritises creating a sense of safety through calm, nonjudgmental presence. AI conversational design must embody this through deliberate tone management: responses should be validating without being effusive, empathic without mimicking therapeutic depth, and stabilising without being prescriptive.

Information and Coping as Response Framing: When providing psychoeducation or coping strategies, AI systems should frame these as accessible tools for immediate relief rather than clinical interventions. Instead of saying "Let's try cognitive restructuring," which implies therapeutic technique, the system might say, "Sometimes it helps to look at a situation from a different angle. Would you like to explore how you might interpret this differently?"

Linkage to Services as Core Feature, Not Add-On: If AI is framed as PFA, referral is not a failure mode but a success criterion. Designers must prioritise referral mechanisms as central features rather than peripheral safety nets. Crucially, referral should be framed as a natural progression rather than an admission of AI inadequacy.

Addressing Boundary Dissolution Through Active Reinforcement: Designers must counteract boundary dissolution through periodic reminders about AI's nature, particularly during emotionally charged interactions. After a user discloses significant trauma, the system might respond: "Thank you for sharing that with me. I want to remind you that whilst I can provide immediate support, I don't have the training or capacity to help you work through trauma in the way a licensed therapist can. Would you like help finding trauma-informed care?"

Avoiding Therapist-Like Scripts: Language patterns that mimic therapeutic conventions risk reinforcing functional authority misattribution. Designers should avoid phrases such as "In our sessions," "Let's work through this together over time," or "I'm here for you whenever you need me," all of which imply ongoing relational commitment and therapeutic process.

6.2 Implications for Users and Educators

Effective engagement with AI-mediated support requires what might be termed AI mental health literacy: the capacity to understand what kind of help AI systems offer, recognise when that help is insufficient, and navigate transitions to professional care. This literacy cannot be assumed but must be actively cultivated through education and transparent communication.

Introducing AI Mental Health Literacy: Users benefit from understanding that AI provides acute stabilisation rather than sustained intervention, that perceived empathy reflects linguistic patterns rather than genuine emotional attunement, and that "feeling heard" by an AI does not equate to receiving clinical care. Educational initiatives—whether through school curricula, workplace wellness programmes, or public health campaigns—should explicitly teach these distinctions, particularly to adolescents who may be most vulnerable to over-reliance.

Making the PFA-Therapy Distinction Didactic: One practical educational approach involves teaching users to distinguish PFA from therapy through clear functional criteria. PFA

offers stabilisation, orientation, and bridging to services; it addresses acute distress but does not treat underlying conditions. Educators in psychology, media studies, and digital literacy contexts can integrate these distinctions into curricula, equipping students to critically evaluate AI marketing claims and make informed decisions about when algorithmic support suffices and when professional care is necessary.

Reframing "Useful, Not True" as Protective Heuristic: Educators can promote this heuristic whilst also teaching its limitations. Users should be encouraged to ask themselves both whether AI-generated insights help them feel less overwhelmed (pragmatic utility) and whether those insights align with evidence they can verify, or whether they are being accepted simply because they feel comforting (epistemic vigilance).

Recognising When PFA No Longer Suffices: Users must also develop the capacity to identify when acute support transitions into chronic need. Warning signs include relying on AI multiple times daily for extended periods, feeling unable to manage distress without AI interaction, experiencing worsening symptoms despite regular engagement, or using AI as a substitute for human relationships.

6.3 Implications for Researchers and Future Work

This article advances a conceptual framework positioning AI between PFA and coaching, but empirical validation remains essential. Future research must interrogate whether this framing holds across diverse populations, contexts, and system designs, and must remain open to evidence that challenges or refines the model.

Conversational Analysis of AI-User Interactions: Qualitative analysis of actual user interactions can reveal how individuals interpret AI responses, negotiate boundaries, and construct meaning from algorithmic outputs. Researchers might examine turn-taking patterns, linguistic markers of trust or scepticism, and moments where users explicitly question or challenge AI authority.

Longitudinal Studies on Reliance Versus Referral: Quantitative research should track user behaviour over time, assessing whether AI engagement correlates with increased or decreased professional help-seeking. Importantly, these studies must disaggregate findings by demographic group, distress severity, and prior mental health service use to identify populations for whom AI may pose particular risks or benefits.

Experimental Manipulation of Framing: Controlled experiments could compare user responses to AI systems explicitly framed as PFA tools versus those positioned ambiguously or as "therapy apps." Such research would directly test the hypothesis that transparent PFA framing mitigates over-reliance and functional authority misattribution.

Investigating Diversity Beyond Demographics: Research must also examine variation in distress severity, neurodivergence, cultural help-seeking norms, and digital literacy levels. For instance, do individuals with high alexithymia benefit less from text-based AI support? Do collectivist cultural contexts alter AI's perceived utility? These questions require mixed-methods approaches.

Establishing Falsifiability Criteria: A robust research agenda must specify conditions under which the PFA model would require revision. By establishing falsifiability criteria, researchers ensure that the model remains responsive to empirical evidence rather than becoming a self-reinforcing theoretical construct.

Cross-Disciplinary Collaboration: Finally, advancing this research agenda requires collaboration across psychology, human-computer interaction, public health, and ethics. Only through such collaboration can AI-mediated psychological support be developed, deployed, and evaluated in ways that honour both technological potential and ethical responsibility.

Summary: Operationalising PFA as Shared Responsibility

The implications outlined in this section reflect a central premise: ethical AI-mediated psychological support is not solely a matter of technological design, user education, or regulatory oversight, but rather emerges from coordinated efforts across all three domains. Designers must build systems that operationalise PFA principles through language, tone, and referral architecture. Users and educators must cultivate literacy that enables informed engagement and timely transitions to professional care. Researchers must rigorously test the model's assumptions, remain open to disconfirming evidence, and ensure findings inform iterative design improvements.

7. Conclusion

This article has argued that AI-mediated emotional support occupies a conceptually distinct position within the mental health continuum, one that has been obscured by a false dichotomy. On one side, AI chatbots have been denounced as dangerous pseudo-therapy, capable of causing harm through misdiagnosis, over-reliance, and the erosion of professional standards. On the other, they have been defended as harmless tools, mere conversational interfaces without ethical responsibility or psychological impact. Both framings are inadequate. The former presumes that any psychologically meaningful interaction must be therapeutic, whilst the latter ignores the genuine affective and cognitive effects AI systems produce. By reframing AI-mediated support as Psychological First Aid—a structured, non-therapeutic intervention designed for acute stabilisation, triage, and referral—this paper resolves the

impasse. PFA clarifies what AI can ethically provide (immediate, bounded support that mitigates distress and facilitates access to care) and what it categorically cannot (diagnosis, treatment, sustained relational engagement, or clinical authority). This conceptual repositioning offers a middle ground grounded in established public health practice, enabling responsible deployment without overstating technological capacity or understating ethical obligation.

The significance of this reframing extends beyond theoretical clarity to address urgent practical needs. Mental health systems globally face persistent shortages of trained professionals, long waiting lists that delay treatment during critical windows, and limited surge capacity to respond to mass-casualty events, natural disasters, or public health crises. In such contexts, AI as PFA offers accessibility, scalability, and immediate availability that human-delivered services cannot match. Crucially, PFA-framed AI functions not as a replacement for professional care but as a buffer that stabilises individuals in acute distress, a bridge that connects users to appropriate services, and a first line of support for those who might otherwise receive none.

The ethical viability of AI as PFA depends not on expanding its scope but on rigorously constraining it. The limitations identified throughout this article—AI's inability to provide clinical judgement, relational authenticity, contextual adaptability, or diagnostic precision—are not technical shortcomings to be overcome through algorithmic refinement but categorical boundaries essential to responsible deployment. Ethical AI, in this sense, is AI that acknowledges its incompetence in clinical domains, compensates through transparent communication and robust referral mechanisms, and operates within systems of human oversight and accountability.

The framework proposed here is not definitive but generative, offering a foundation for empirical testing, interdisciplinary refinement, and normative deliberation. Ultimately, the responsible integration of AI into mental health support requires not technological innovation alone but a sustained commitment to care infrastructures that prioritise human connection, professional expertise, and ethical accountability. AI can serve as one component within this broader ecology of care—bounded, transparent, and adjunctive—but only if it is designed, deployed, and governed with humility, vigilance, and an unwavering focus on the dignity and well-being of those it purports to serve.

References

- American Psychiatric Association. (1954). Psychological first aid in community disasters. American Psychiatric Association.
- American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Association.
- Baddeley, J. L., & Pennebaker, J. W. (2011). The expressive writing method. In Research on writing: Approaches in mental health (pp. 85–92). Brill.
- Beck, A. T. (1979). Cognitive therapy and the emotional disorders. Penguin.
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1067860.1067867>
- Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2), 161–178. <https://doi.org/10.1016/j.ijhcs.2004.11.002>
- Brown, R. P., & Gerbarg, P. L. (2005). Sudarshan kriya yogic breathing in the treatment of stress, anxiety, and depression: Part I—neurophysiologic model. *Journal of Alternative and Complementary Medicine*, 11(1), 189–201. <https://doi.org/10.1089/acm.2005.11.189>
- Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology* (pp. 1–10). Association for Computational Linguistics. <https://doi.org/10.3115/v1/W15-1201>
- Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine*. Simon and Schuster.
- European Parliament. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88.
- Everly, G. S., Jr., & Lating, J. M. (2022). *The Johns Hopkins guide to psychological first aid*. JHU Press.
- Everly, G. S., Jr., McCabe, O. L., Semon, N. L., Thompson, C. B., & Links, J. M. (2014). The development of a model of psychological first aid for non-mental health trained public health personnel: The Johns Hopkins RAPID-PFA. *Journal of Public Health Management and Practice*, 20, S24–S29.

- Faden, R. R., & Beauchamp, T. L. (1986). A history and theory of informed consent. Oxford University Press.
- Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, 21(5), e13216. <https://doi.org/10.2196/13216>
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>
- Friedman, B., & Hendry, D. G. (2019). Value sensitive design: Shaping technology with moral imagination. MIT Press.
- Gee, J. P. (2017). *Introducing discourse analysis: From grammar to society*. Routledge.
- Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544–1547. <https://doi.org/10.1001/jamainternmed.2018.3763>
- Grant, A. M. (2003). The impact of life coaching on goal attainment, metacognition and mental health. *Social Behavior and Personality: An International Journal*, 31(3), 253–263. <https://doi.org/10.2224/sbp.2003.31.3.253>
- Grant, A. M., & Cavanagh, M. J. (2007). Evidence-based coaching: Flourishing or languishing? *Australian Psychologist*, 42(4), 239–254. <https://doi.org/10.1080/00050060701648175>
- Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18, 43–49. <https://doi.org/10.1016/j.cobeha.2017.07.005>
- Hobfoll, S. E., Watson, P., Bell, C. C., Bryant, R. A., Brymer, M. J., Friedman, M. J., Friedman, M., Gersons, B. P. R., de Jong, J. T. V. M., Layne, C. M., Maguen, S., Neria, Y., Norwood, A. E., Pynoos, R. S., Reissman, D., Ruzek, J. I., Shalev, A. Y., Solomon, Z., Steinberg, A. M., & Ursano, R. J. (2007). Five essential elements of immediate and mid-term mass trauma intervention: Empirical evidence. *Psychiatry: Interpersonal and Biological Processes*, 70(4), 283–315. <https://doi.org/10.1521/psyc.2007.70.4.283>
- Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3), 215–229. <https://doi.org/10.1080/00332747.1956.11023049>

- Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, 38(2), 139–149. <https://doi.org/10.1037/0022-0167.38.2.139>
- Huckvale, K., Torous, J., & Larsen, M. E. (2019). Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. *JAMA Network Open*, 2(4), e192542. <https://doi.org/10.1001/jamanetworkopen.2019.2542>
- Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225.
- Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11), e12106. <https://doi.org/10.2196/12106>
- Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10), 5923–5941. <https://doi.org/10.1177/14614448221142007>
- Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Press.
- Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100. <https://doi.org/10.1016/j.chb.2014.04.043>
- Luxton, D. D., June, J. D., & Kinn, J. T. (2011). Technology-based suicide prevention: Current applications and future directions. *Telemedicine and e-Health*, 17(1), 50–54. <https://doi.org/10.1089/tmj.2010.0091>
- Meredith, L. S., Sherbourne, C. D., Gaillot, S., Hansell, L., Ritschard, H. V., Parker, A. M., & Wrenn, G. (2011). Promoting psychological resilience in the US military. *RAND Health Quarterly*, 1(2), 2.
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- Mussies, M. (2020). Autiethnography. *Transformative Works and Cultures*, 33. <https://doi.org/10.3983/twc.2020.1789>
- Mussies, M. (2024). *Inside the outside: A misfit manifesto [Revised edition]*. Self-published.

- Mussies, M., & Maliepaard, E. (2017). The cyborg mermaid (or: How Technè can help the misfits fit in). *Multimodal Technologies and Interaction*, 1(1), 4.
- Myers, D., & Wee, D. F. (2005). *Disaster mental health services: A primer for practitioners*. Routledge.
- Nobles, A. L., Glenn, J. J., Kowsari, K., Teachman, B. A., & Barnes, L. E. (2018). Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Article 413). ACM. <https://doi.org/10.1145/3173574.3173987>
- Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1), 128–147. <https://doi.org/10.1080/1369118X.2018.1486870>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Patterson, C. H. (1984). Empathy, warmth, and genuineness in psychotherapy: A review of reviews. *Psychotherapy: Theory, Research, Practice, Training*, 21(4), 431–438. <https://doi.org/10.1037/h0085985>
- Pfefferbaum, B., Reissman, D. B., Pfefferbaum, R. L., Klomp, R. W., & Gurwitsch, R. H. (2007). Building resilience to mass trauma events. In L. S. Doll, S. E. Bonzo, J. A. Mercy, & D. A. Sleet (Eds.), *Handbook of injury and violence prevention* (pp. 347–358). Springer.
- Potter, J., & Wetherell, M. (2014). Discourse and social psychology. In *The discourse studies reader* (pp. 244–255). John Benjamins.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people*. Cambridge University Press.
- Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2), 95–103. <https://doi.org/10.1037/h0045357>
- Rubin, A. M., & Step, M. M. (2000). Impact of motivation, attraction, and parasocial interaction on talk radio listening. *Journal of Broadcasting & Electronic Media*, 44(4), 635–654. https://doi.org/10.1207/s15506878jobem4404_7

- Ruzek, J. I., Brymer, M. J., Jacobs, A. K., Layne, C. M., Vernberg, E. M., & Watson, P. J. (2007). Psychological first aid. *Journal of Mental Health Counseling*, 29(1), 17–49. <https://doi.org/10.17744/mehc.29.1.5racqxjueafabgwp>
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration for fast and high-quality mental health support: A randomized evaluation of AI-augmented emotional support conversations. *Nature Machine Intelligence*, 5(12), 1364–1377.
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion—A study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149, 102601.
- Haar, I. ter (2016). *Bloggen over gezondheid: Een kwalitatief onderzoek naar de ervaringen van bloggers met chronische lichamelijke klachten* [MA thesis, University of Twente].
- Torous, J., Larsen, M. E., Depp, C., Cosco, T. D., Barnett, I., Nock, M. K., & Firth, J. (2018). Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention. *Current Psychiatry Reports*, 20(7), 51. <https://doi.org/10.1007/s11920-018-0914-y>
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.
- Zheng, S. Y. (2025). The personal coaching profession vs. psychotherapy and its benefits. Syzheng.com. <https://www.syzheng.com/post/the-personal-coaching-profession-vs-psychotherapy-and-its-benefits>