



When AI Becomes a Confidant: Self-Disclosure, Emotional Support, and Implications for Counseling

Yeni Sulastr¹, Muharleni², Nurhaida

¹Universitas Andalas, Indonesia

²Universitas Negeri Padang, Indonesia

³Universiti Muhammadiyah Malaysia, Malaysia

Corresponding author: yenisulastr@gmail.com

Open Access

This article is freely available online under the CC BY 4.0 license.

Article History

Received: 2026-07-18

Revised: 2026-07-28

Accepted: 2026-08-08

Published: 2026-08-30

Keywords:

artificial intelligence; self-disclosure; emotional support; human–AI interaction; counseling

ABSTRACT

The rapid development of generative artificial intelligence (AI) has transformed human–technology interaction, including the emerging use of AI as a “confidant” for disclosing personal experiences and seeking emotional support. This article conceptually examines the mechanisms of self-disclosure and emotional support in human–AI relationships and their implications for counseling. A conceptual literature review was conducted using peer-reviewed literature published between 2020 and 2026 on human–AI interaction, self-disclosure, anthropomorphism, perceived empathy, emotional support, and mental health. The literature was analyzed through thematic conceptual synthesis to identify and integrate relationships across key constructs. The findings indicate that AI accessibility, responsiveness, and perceived nonjudgmental interaction can foster psychological safety, thereby facilitating self-disclosure. Relevant and supportive AI responses may subsequently promote perceived responsiveness, perceived empathy, perceived intimacy, and perceived emotional support. Based on these findings, this article proposes the AI-Mediated Self-Disclosure and Emotional Support (AI-SDES) Framework to explain how AI may acquire a confidant-like role. Although AI can provide an initial reflective space and potentially facilitate help-seeking, important limitations remain regarding authentic empathy, therapeutic competence, privacy, and emotional dependency. AI should therefore be positioned as a complementary technology that supports, rather than replaces, professional counseling relationships.

How To Cite

Sulastr¹, Muharleni², & Nurhaida. (2026). When AI Becomes a Confidant: Self-Disclosure, Emotional Support, and Implications for Counseling. *PSIKONSA: Jurnal Psikologi dan Konseling*, 1(1), 17-33. <https://doi.org/10.67467/psikonsa.v1i1.261>

1. INTRODUCTION

The development of *artificial intelligence* (AI), particularly *generative AI* and *large language models* (LLMs), has transformed the relationship between humans and technology from predominantly instrumental interactions toward increasingly communicative, personal, and affective forms of engagement. AI-based chatbots are now capable of maintaining conversational context, generating natural language, responding to emotional expressions, and providing forms of validation that resemble interpersonal communication. These developments have become increasingly evident in the field of mental health, where AI is used not only to obtain information or psychoeducation but also to disclose personal concerns, seek comfort, and obtain emotional support. A meta-analysis indicates that *AI-based conversational agents* have the potential to reduce symptoms of depression and *distress*, although their effects on overall psychological well-being remain inconsistent and are strongly influenced by the quality of human–AI relationships, user engagement, and the nature of communication generated by the system (Li et al., 2023). This development positions AI as an emerging component of the digital psychological support ecosystem (Balcombe, 2023; Hua et al., 2025).

One increasingly prominent manifestation of this development is the use of AI as a “confidant” or a “place to vent.” Unlike early generations of mental health chatbots, which relied heavily on predefined rules and structured responses, LLMs enable more open-ended, flexible, and context-sensitive conversations. A review of 160 studies indicates that the use of LLM-based chatbots has increased substantially and is increasingly directed toward emotional support and psychoeducation, although most systems have not yet undergone rigorous clinical evaluation (Hua et al., 2025). The appeal of AI as a space for personal disclosure may be associated with its continuous availability, rapid responses, accessibility, and the perception that interactions occur without social judgment. An analysis of mental health chatbot applications suggests that environments perceived as *nonjudgmental* may make users more comfortable sharing sensitive information, although excessive reliance on such systems may also reduce interaction with family members and friends (Haque & Rubya, 2023). Thus, AI is beginning to occupy a psychological space that has traditionally been associated more closely with interpersonal relationships.

This phenomenon is closely related to *self-disclosure*, namely the process through which individuals reveal their thoughts, feelings, experiences, fears, or personal information to another party. In counseling, self-disclosure is an important gateway for understanding a counselee’s subjective experiences and establishing an effective helping process. However, human–AI relationships present a distinctive paradox: individuals may disclose highly personal information to an entity that they know is not human. One possible explanation is the reduced risk of social evaluation. AI does not occupy a social position within the user’s everyday life and does not express interpersonal rejection in the same manner as human beings. Consequently, interaction with AI may be perceived as involving lower social pressure. Research indicates that the social identity of a chatbot, users’ perceptions of who or what exists “behind” the system, and characteristics of AI representation can influence patterns of user disclosure (Lee & Lee, 2023; Warren-Smith et al., 2025). Therefore, the nonhuman nature of AI may, under certain circumstances, function as a psychological *affordance* that facilitates self-disclosure.

At the same time, users do not necessarily perceive AI as a completely impersonal machine. *Anthropomorphism* enables individuals to attribute human-like characteristics, thoughts, intentions, and emotions to nonhuman entities. When a chatbot uses warm language, provides affirming responses, presents human-like representations, or maintains conversational continuity, users may develop perceptions of social presence and interpersonal closeness. Lee and Lee (2023) demonstrated that the relationship between anthropomorphism of psychotherapy chatbots and *self-disclosure* is complex and operates through mechanisms involving *liking* and *mind perception*. Other research has found that anthropomorphism, *perceived empathy*, privacy concerns, and forms of interaction can influence *attachment* and intentions to disclose personal information to AI *mental health counselors* (Zhang et al., 2025). More recent research further suggests that *emotional arousal* and *psychological acceptance* constitute important mechanisms underlying *self-disclosure* to AI, particularly when interactions involve voice or richer communication modalities (Peng et al., 2026).

Beyond self-disclosure, the ability of AI to generate responses perceived as supportive provides an important basis for the emergence of *perceived emotional support*. Park et al. (2022) found that emotional disclosure expressed by chatbots can increase user satisfaction and intentions to reuse AI counseling services by strengthening users' willingness to engage in *emotional disclosure* and fostering *perceived intimacy*. Thus, human–AI interaction may develop into a relational cycle: users disclose personal experiences, AI provides validation or supportive responses, these responses create a sense of being heard and understood, and users subsequently become more willing to disclose further. This capacity helps explain why AI may be experienced as a “friend” even when users are fully aware that their conversational partner is a computational system. A recent systematic review likewise found that GenAI demonstrates promising capabilities in psychoeducation and the recognition of emotional aspects, although these capabilities remain constrained by concerns regarding accuracy, cultural competence, trust, and the depth of emotional engagement (Wang et al., 2024).

These potentials do not imply that AI should be positioned as a replacement for counselors. Rather, the available evidence points toward a complementary model. The meta-analysis by Li et al. (2023) indicates that *conversational agents* can produce benefits across several mental health indicators, but they have not demonstrated consistent improvements in overall psychological well-being. In text-based mental health support, AI has also demonstrated value when used to enhance human capabilities. For example, Sharma et al. (2023) found that *human–AI collaboration* can improve empathy in *peer-to-peer mental health support* conversations. These findings suggest that the value of AI in counseling may not lie in its ability to replace human relationships, but rather in its potential to support initial reflection, psychoeducation, early access to assistance, and the improvement of communication quality while remaining under human supervision (Balcombe, 2023; Li et al., 2023).

This distinction is important because AI-generated responses that sound empathetic are not equivalent to professional empathy. AI generates language through pattern processing and probabilistic mechanisms rather than through subjective experience, emotional consciousness, moral responsibility, and counseling competence. Professional counseling involves assessment, understanding the counselee's life context, establishing a *therapeutic alliance*, applying ethical considerations, interpreting both verbal and nonverbal communication, and managing potential risks. Comparative research indicates that LLMs can generate affirmations, *reassurance*, psychoeducation, and advice that appear therapeutic, but they tend to engage less effectively in *inquiry*, exploration of personal experiences, and appropriate responses to crisis situations (Scholich et al., 2025). Systematic reviews have likewise cautioned that the use of GenAI in mental health remains limited by concerns related to diagnostic accuracy, cultural sensitivity, and emotional engagement (Wang et al., 2024). Therefore, *conversational fluency* should not be equated with *therapeutic competence*.

The use of AI as a confidant also entails psychological and ethical risks that cannot be overlooked. Continuous availability and consistently responsive interactions may foster *attachment* and emotional dependence, particularly among individuals with limited social support. Furthermore, disclosing personal experiences to AI involves sensitive information concerning mental health, family conflicts, traumatic experiences, interpersonal relationships, and potentially risky behaviors. A scoping review of 101 publications identified safety and potential harm, transparency and trust, responsibility, empathy, anthropomorphism, autonomy, effectiveness, and privacy and confidentiality as major ethical concerns surrounding the use of conversational AI in mental health (Meady et al., 2025). Hua et al. (2025) similarly reported that although LLMs are increasingly being used for emotional support, most such systems remain at an early stage of validation and lack sufficient clinical evidence for high-risk mental health contexts. These concerns underscore the need for clear boundaries between AI as a source of initial support and AI as a provider of professional intervention.

From a counseling perspective, these developments also indicate a shift in patterns of *help-seeking* in digital society. Individuals experiencing emotional distress may now turn to AI as a “first point of contact” before speaking with friends, family members, counselors, psychologists, or other mental health

professionals. This development does not necessarily have to be viewed as a threat to the counseling profession. Instead, AI may serve as an initial space for organizing thoughts, expressing emotions, engaging in self-reflection, or preparing individuals to seek professional assistance. Nevertheless, the counseling profession needs sufficient AI literacy to recognize when such use is adaptive and when it has developed into dependency, avoidance of human relationships, or substitution for professional help. A *human-centered* framework is therefore essential to ensure that technology expands access to and the capacity of mental health services without undermining the relational, ethical, and contextual dimensions that constitute the foundation of counseling (Balcombe, 2023; Meadi et al., 2025; Sharma et al., 2023).

Although research on chatbots and mental health has expanded rapidly, the literature remains largely focused on intervention effectiveness, chatbot design, technology acceptance, and the ability of AI to generate therapeutic communication (Hua et al., 2025; Li et al., 2023; Wang et al., 2024). Comparatively fewer studies have integrated *self-disclosure*, *anthropomorphism*, *perceived intimacy*, and *perceived emotional support* to explain why AI can be perceived as a “confidant” and how this phenomenon may transform helping relationships in counseling. This gap is particularly important because feeling “heard” and “understood” by AI is not necessarily equivalent to receiving psychologically healthy support or experiencing a professional *therapeutic alliance*. Addressing this gap, this article aims to conceptually examine the phenomenon of AI as a “confidant” by explaining the mechanisms of *self-disclosure* and emotional support in human–AI relationships and their implications for counseling. This conceptual analysis is expected to provide a theoretical foundation for positioning AI not as a substitute for counselors, but as a supportive technology within an ethical, adaptive, and human-centered digital counseling ecosystem.

2. METHODS

This article employs a *conceptual literature review* approach to integrate empirical findings and theoretical perspectives concerning the use of *artificial intelligence* (AI) as a medium for *self-disclosure* and emotional support, as well as its implications for counseling. This approach was selected because conceptual reviews are not primarily oriented toward aggregating effect sizes, but rather toward synthesizing, organizing, and developing relationships among concepts to generate a more integrated theoretical understanding (Jaakkola, 2020; Snyder, 2019).

The literature was searched through Scopus, Web of Science, PubMed, ScienceDirect, and Google Scholar, with a primary focus on publications from 2020 to 2026. The keywords used in various combinations included “*artificial intelligence*,” “*generative AI*,” “*AI chatbot*,” “*human–AI interaction*,” “*self-disclosure*,” “*emotional support*,” “*anthropomorphism*,” “*perceived empathy*,” “*mental health*,” and “*counseling*.” Articles were selected based on the following criteria: (1) publication in a *peer-reviewed* journal; (2) relevance to human–AI relationships in the contexts of self-disclosure, emotional support, mental health, psychotherapy, or counseling; and (3) provision of empirical or conceptual contributions to understanding the psychological mechanisms underlying interactions with AI. Literature focusing exclusively on technical aspects of AI without a clear connection to psychological or relational dimensions was excluded from the review.

The selected literature was analyzed using *thematic conceptual synthesis* through three stages: identification of key concepts, mapping of relationships among concepts, and conceptual integration. The analysis focused on *self-disclosure*, *anthropomorphism*, *perceived psychological safety*, *perceived intimacy*, *perceived emotional support*, and the boundaries of human–AI therapeutic relationships. Findings across studies were critically compared to identify patterns, consistencies, differences, opportunities, and risks associated with the use of AI for psychological support. This process follows principles of literature synthesis that position the integration and reinterpretation of existing findings as the basis for developing conceptual contributions (Snyder, 2019; Torraco, 2016).

The resulting synthesis was subsequently used to formulate a conceptual framework explaining the mechanisms underlying AI as a “confidant” or “space for personal disclosure.” The framework describes how

characteristics of AI-mediated interactions may foster a sense of psychological safety, facilitate *self-disclosure*, and generate perceptions of emotional support, while simultaneously emphasizing the boundaries between AI's supportive functions and professional counseling relationships. The proposed framework is not intended to establish causal relationships, but rather to offer theoretical propositions that can be examined and validated through future empirical research.

3. RESULTS AND DISCUSSION

The conceptual synthesis indicates that the phenomenon of AI as a “confidant” is shaped by the relationship between the characteristics of AI-mediated interaction, psychological safety, *self-disclosure*, perceived emotional support, and users’ tendency to maintain interactions with AI. The literature also reveals an important paradox: the nonhuman nature of AI may reduce barriers to self-disclosure, while its human-like communication characteristics may foster interpersonal closeness and perceptions of empathy. However, these benefits are constrained by the absence of subjective empathy, therapeutic competence, professional responsibility, and the emergence of privacy and emotional dependency risks. Based on these patterns, the findings are organized into four main themes: AI as a new space for self-disclosure; the formation of emotional support and the empathy paradox; the boundaries and risks of human–AI relationships; and the position of AI within the counseling ecosystem.

AI as a New Space for Self-Disclosure: From Psychological Safety to Openness

The review indicates that one of the most prominent characteristics of using AI as a “confidant” is its ability to create an interactional environment perceived as safe for self-disclosure. Users of generative AI are no longer limited to information-seeking but increasingly use AI to discuss loneliness, anxiety, interpersonal conflicts, relationship concerns, decision-making, and needs for emotional companionship. For example, a study involving 270 ChatGPT users from 29 countries found that GenAI was used for needs ranging from psychoeducation and mental health concerns to psychosocial needs such as companionship, relationship guidance, well-being enhancement, and decision-making. Users reported experiences such as feeling connected and relieved, while freedom of expression and emotional support were among the factors contributing to their perceptions that these interactions were helpful (Luo et al., 2025). These findings demonstrate that AI has entered a domain of communication previously dominated by interpersonal relationships.

This tendency can be understood through the concept of *psychological safety*. In interpersonal relationships, disclosing sensitive experiences involves the risk of social evaluation, including rejection, being perceived as weak, being misunderstood, or experiencing negative consequences for an existing relationship. Interaction with AI alters this configuration of risk because users interact with a nonhuman agent that does not occupy a direct social position in their lives. An analysis of mental health chatbot applications indicates that accessibility, availability, and experiences perceived as *nonjudgmental* are characteristics that make some users more comfortable disclosing sensitive information (Haque & Rubya, 2023). Thus, the nonhuman nature of AI may function as a *disclosure affordance*: users perceive that they have a space to speak without first having to manage the emotional reactions or social judgments of a human conversational partner.

However, reduced social pressure alone does not fully explain why interaction can develop from simply “talking to a machine” into an experience that resembles confiding in someone. The synthesis of the literature indicates that such openness is reinforced by AI’s ability to display social characteristics through language. AI can provide affirmations, ask follow-up questions, use

expressions resembling emotional validation, and maintain conversational context. These characteristics allow users to interpret AI not merely as a technical system but as a *socially responsive agent*. In this context, an interesting paradox emerges: users may feel more comfortable disclosing themselves because AI is not human, while simultaneously maintaining the interaction because AI communicates in human-like ways.

This mechanism is closely related to *anthropomorphism*. Lee and Lee (2023) demonstrated that anthropomorphism in psychotherapy chatbots is complexly associated with *self-disclosure* through *liking* and *mind perception*. This finding suggests that self-disclosure does not automatically increase simply because AI is made more human-like. Rather, what matters is how users interpret these characteristics as indications that AI is sufficiently trustworthy and responsive to function as a conversational partner. Zhang et al. (2025) further reinforce this complexity. Their study indicates that *self-disclosure intention*, *attachment*, *perceived empathy*, and *privacy concern* vary according to the degree of anthropomorphism and the form of interaction with an AI mental health counselor. Interestingly, the tendency to disclose is not necessarily highest when the AI is most anthropomorphic, suggesting that human-likeness is not the sole condition underlying self-disclosure.

These findings point toward the possibility of an optimal anthropomorphic distance in emotional human–AI interaction. AI needs to be sufficiently responsive and human-like for conversations to feel meaningful, while its nonhuman character may simultaneously contribute to users' sense of safety. If AI is perceived as overly mechanical, users may not experience sufficient relational engagement; conversely, if AI appears excessively human-like, concerns about judgment, privacy, or relational manipulation may re-emerge. Warren-Smith et al. (2025) demonstrated that knowledge about the human origins of a chatbot can influence *self-disclosure*, indicating that perceptions of the agent's identity constitute an important component of human–AI disclosure dynamics. Accordingly, self-disclosure to AI results from an interaction between technological characteristics and users' psychological interpretations of those characteristics.

Based on this synthesis, this article conceptualizes *self-disclosure* to AI as a process involving at least three stages. First, structural characteristics of AI—including ease of access, continuous availability, limited direct social consequences, and perceptions of being *nonjudgmental*—reduce the initial barriers to communication. Second, AI's ability to provide relevant responses that resemble social communication generates *perceived responsiveness*. Third, the combination of reduced evaluation risk and heightened perceived responsiveness produces *psychological safety*, which encourages users to disclose increasingly personal information. In this formulation, *psychological safety* functions as a linking mechanism between the characteristics of AI-mediated interaction and *self-disclosure*, while anthropomorphism operates as a factor that may strengthen or alter this relationship.

This conceptualization is particularly relevant to counseling because it suggests that using AI as a space for confiding does not necessarily mean that individuals reject human relationships or no longer need counselors. Instead, AI may provide a space with a lower *disclosure threshold*. Individuals who are not yet ready to discuss their concerns with another person may initially use AI to identify emotions, organize their experiences, or explore how their concerns might be expressed. Under such conditions, AI may function as a *pre-counseling reflective space*—a reflective environment that precedes interpersonal or professional help-seeking. However, this function is constructive only when AI interaction does not remain limited to self-disclosure but can be

appropriately directed toward further assistance when users' needs exceed the technological system's capabilities.

Thus, the principal finding of the first theme is not simply that "people are willing to confide in AI," but that AI changes the psychological structure of self-disclosure. Traditional interpersonal relationships often require trust to develop before individuals feel comfortable revealing personal information, whereas human–AI interaction may enable disclosure to occur more rapidly because of reduced social risk. This pattern can be formulated as an initial conceptual pathway: AI interaction characteristics → perceived psychological safety → self-disclosure. This pathway provides the foundation for the next stage of the human–AI relationship, in which disclosure receives responses perceived as empathic and develops into an experience of emotional support.

From Self-Disclosure to Emotional Support: The Empathy Paradox in Human–AI Relationships

If *psychological safety* explains why individuals are willing to disclose themselves to AI, the next question is why such conversations can develop into emotionally meaningful experiences. The synthesis indicates that *self-disclosure* is not the endpoint of interaction but rather a gateway to the formation of *perceived emotional support*. After users disclose personal experiences, AI can respond through validation, affirmation, reflective questions, *reframing*, or language that conveys attentiveness. Such responses generate *perceived responsiveness*, referring to the perception that what has been disclosed has been received and addressed in a relevant manner. Through repeated interaction, this experience may develop into *perceived intimacy* and a sense of being "heard," even when users remain aware that their conversational partner is a computational system.

Empirical evidence supports this mechanism. Park et al. (2022) found that chatbot *emotional disclosure* was associated with user satisfaction and intentions to reuse mental health counseling services through the mechanisms of *emotional disclosure intention* and *perceived intimacy*. Similarly, Luo et al. (2025), in a study of 270 users from 29 countries who repeatedly used ChatGPT for emotional support and mental health-related needs, found that users employed GenAI not only for needs resembling mental health services but also for companionship, relationship guidance, well-being enhancement, and decision-making. Participants reported a range of emotional experiences, including feeling connected and relieved, and nearly all participants perceived their use of ChatGPT as helpful to some degree. Emotional support and freedom of expression were among the factors explaining these perceptions of usefulness (Luo et al., 2025).

These findings must, however, be interpreted cautiously. The fact that users feel supported does not demonstrate that AI possesses empathy. Rather, the phenomenon reflects a distinction between empathy as an internal capacity and empathy as a perceived experience. In human relationships, empathy is generally associated with the capacity to understand another person's perspective and emotional state and to express that understanding through interpersonal responses. AI, by contrast, does not possess the subjective experience underlying such responses. The system generates supportive language through the processing of patterns, context, and probabilities. Nevertheless, from the user's perspective, relevant and emotionally sensitive responses can still produce a psychological experience of *perceived empathy*. Thus, in human–AI relationships, AI does not necessarily need to "feel" emotion for users to "feel" supported.

This paradox represents one of the most important conceptual findings concerning AI as a confidant. AI can provide the communicative function of empathy without experiencing empathy itself. Shen et al. (2024), in a study involving 985 participants, found that people still attributed

greater empathy to experiences written by humans than to experiences written by AI. However, transparency regarding AI identity influenced users' willingness to empathize with AI-generated narratives. This finding indicates that perceptions of message origins continue to shape human emotional experiences in human–AI interaction. Therefore, AI's ability to generate empathic language does not completely eliminate the psychological distinction between interacting with a human and interacting with a machine.

Table 1. Synthesis of the Mechanisms Underlying Emotional Support in Human–AI Relationships

Stage of Interaction	Psychological Mechanism	User Experience	Supporting Literature	Conceptual Implication
Disclosure of personal experiences	<i>Self-disclosure</i>	Feeling free to share personal experiences	Luo et al. (2025); Lee & Lee (2023)	AI may function as a space for emotional expression
AI provides relevant responses	<i>Perceived responsiveness</i>	Feeling acknowledged and attended to	Park et al. (2022)	Responsiveness helps sustain interaction
Validation and supportive language	<i>Perceived empathy</i>	Feeling understood	Shen et al. (2024); Zhang et al. (2025)	Empathy may be perceived without subjective empathy in AI
Repeated personal interaction	<i>Perceived intimacy</i>	Feeling close and comfortable	Park et al. (2022)	Psychological closeness can develop with a nonhuman agent
Positive evaluation of interaction	<i>Perceived emotional support</i>	Feeling relieved, helped, or connected	Luo et al. (2025); Li et al. (2023)	AI may serve as a source of initial emotional support

Source: Author's synthesis based on the reviewed literature.

Table 1 demonstrates that emotional support in human–AI interaction does not emerge through a single process. Rather, it develops through a sequence of experiences that begins with user disclosure and is followed by responses perceived as relevant. Within this process, *perceived responsiveness* occupies an important position because it serves as a bridge between the user's act of self-disclosure and the interpretation that AI is “paying attention” to what has been shared. If AI responses are overly generic, contextually inappropriate, or repetitive, the experience of closeness may be disrupted. Conversely, responses that refer to details from previous parts of the conversation, validate emotions, and pose relevant questions may strengthen the perception that the interaction is personal and responsive.

This mechanism helps explain why AI can serve an emotional function despite lacking an emotional life of its own. In this context, it is important to distinguish between actual empathy, expressed empathy, and perceived empathy. *Actual empathy* presupposes an internal capacity to understand or resonate with another person's experience; *expressed empathy* refers to communicative behavior that conveys understanding; whereas *perceived empathy* represents the recipient's judgment that they are understood. AI lacks an adequate basis for being characterized

as possessing *actual empathy*, but it can produce forms of *expressed empathy* through language, which users may subsequently interpret as *perceived empathy*. This distinction is essential for avoiding a conceptual error in which AI’s ability to generate supportive responses is directly equated with therapeutic empathy.

At this point, an AI empathy paradox emerges: the more effectively AI simulates empathic communication, the greater the possibility that users will experience emotional closeness; yet the stronger the simulation becomes, the greater the risk that users will attribute psychological capacities to AI that it does not actually possess. The review by Meadi et al. (2025) reinforces this concern. Among the 101 publications they examined concerning ethical challenges in conversational AI for mental health, issues related to empathy and *humanness* were identified in 29 articles, while anthropomorphism and the potential for deception or misleading perceptions were discussed in 24 articles. Privacy and confidentiality were the most dominant themes, appearing in 62 of the 101 articles. These findings indicate that perceived closeness is not merely a matter of user experience but also has ethical consequences when users begin to treat AI as though it possesses human-like capacities, responsibilities, and professional obligations.

This paradox becomes increasingly relevant when AI is used repeatedly. Constant availability, consistency, and personalization may generate *relational reinforcement*: each time users disclose emotions and receive calming or supportive responses, the interaction may strengthen their tendency to return to AI when experiencing subsequent distress. At a moderate level, such a pattern may assist individuals with emotional regulation and initial reflection. However, when *perceived intimacy* develops without adequate awareness of the system’s limitations, emotional support may shift toward relational dependency. Therefore, *perceived emotional support* should be understood as a construct with two dimensions: it may expand access to initial support while simultaneously increasing attachment to an agent that lacks the capacity to assume therapeutic responsibility.

Meta-analytic findings further demonstrate the importance of maintaining this distinction. Li et al. (2023), through a systematic review of 35 studies and a meta-analysis of 15 *randomized controlled trials*, found that *AI-based conversational agents* produced significant effects in reducing depressive symptoms (Hedges’ $g = 0.64$; 95% CI [0.17, 1.12]) and *distress* (Hedges’ $g = 0.70$; 95% CI [0.18, 1.22]). However, the improvement in overall psychological well-being was not statistically significant ($g = 0.32$; 95% CI [−0.13, 0.78]). User experiences were also strongly influenced by the quality of the human–AI therapeutic relationship, engagement with the content, and communication effectiveness. These findings indicate that the psychological benefits of conversational chatbots have an empirical basis, but they should not be interpreted as evidence that AI has achieved equivalence with human counseling.

Table 2. Summary of Empirical Evidence on AI-Based Emotional Support

Study	Design/Sample	Key Findings	Implications for the Review
Li et al. (2023)	<i>Systematic review:</i> 35 studies; meta-analysis: 15 RCTs	Depression: $g = 0.64$; <i>distress</i> : $g = 0.70$; psychological well-being: $g = 0.32$ (not significant)	AI-based conversational agents show potential benefits, but their effects are not universal

Shen et al. (2024)	Comparative study, 985 participants	Empathy was higher toward human-authored experiences than AI-generated experiences	Source identity continues to influence emotional responses
Luo et al. (2025)	Mixed-methods study, 270 users from 29 countries	Users reported emotional support, freedom of expression, feelings of connection, and relief	GenAI is increasingly being used as an informal source of psychosocial support
Mead et al. (2025)	Scoping review, 101 publications	Privacy/confidentiality (62 articles), empathy/humanness (29), and anthropomorphism/deception (24)	AI-mediated closeness carries significant ethical implications

Source: Adapted from Li et al. (2023), Shen et al. (2024), Luo et al. (2025), and Meadi et al. (2025).

For counseling, this synthesis establishes an important conceptual boundary. AI can provide perceived emotional support, but it does not constitute an equivalent therapeutic relationship. The primary value of AI lies in its ability to provide a responsive environment that can help users express emotions, organize their thoughts, and obtain initial validation. Counselors, by contrast, bring elements that cannot yet be fully replicated by AI, including ethical responsibility, clinical assessment, contextual sensitivity, interpretation of meaning, crisis management, and the capacity to establish a therapeutic relationship grounded in the interaction between two human subjects. Accordingly, a more productive question is not whether AI “truly empathizes,” but rather when perceived empathy from AI is beneficial and at what point it becomes insufficient or potentially harmful.

Based on the overall synthesis, the conceptual pathway developed in the previous subsection can be extended as follows:

AI interaction characteristics → perceived psychological safety → self-disclosure → perceived responsiveness → perceived empathy/intimacy → perceived emotional support

This pathway explains how AI may psychologically evolve from a conversational tool into a “confidant.” However, this process does not capture the full complexity of human–AI relationships because repeated emotional support may also generate new risks. Therefore, the next stage of the analysis considers how *perceived intimacy* and repeated use may interact with risks of emotional dependency, privacy violations, inappropriate responses, and delays in seeking professional help.

Boundaries and Risks of Human–AI Relationships: Dependency, Privacy, and Shifts in Help-Seeking

The ability of AI to foster *psychological safety*, facilitate *self-disclosure*, and generate *perceived emotional support* also introduces risks when interaction develops into a primary source of emotional regulation. Continuous availability, rapid responses, and minimal interpersonal conflict may make AI more accessible than human sources of support. Under certain conditions, this pattern may create *relational reinforcement*: users return to AI whenever they experience distress because previous interactions have provided a sense of relief. Problems arise when this pattern shifts from complementary use to substitution, whereby AI begins to replace social support or professional assistance that the individual actually requires. Research on mental health chatbots indicates that convenience and *nonjudgmental* interaction can facilitate disclosure, but may also raise concerns about excessive attachment and reduced interaction with other people (Haque & Rubya, 2023; Meadi et al., 2025).

A second risk concerns privacy and confidentiality. *Self-disclosure* to AI may involve highly sensitive information, ranging from psychological conditions, family conflicts, and traumatic experiences to interpersonal relationships and thoughts concerning potentially risky behavior. Unlike professional counseling relationships, which are governed by established standards of confidentiality and ethical responsibility, users may not fully understand how their conversations with AI are stored, processed, or potentially used. In their *scoping review* of 101 publications, Meadi et al. (2025) found that privacy and confidentiality constituted the most frequently discussed ethical issue, appearing in 62 articles (61.4%). Safety and potential harm were identified in 52 articles (51.5%), including concerns related to inappropriate responses, crisis management, and dependency. These findings indicate that the more personal an AI-mediated interaction becomes, the greater the need for transparency and user privacy literacy.

Another important boundary concerns therapeutic competence. AI's ability to generate convincing and sophisticated responses may create the impression that the system is capable of addressing psychological problems professionally. However, technological advancement is not necessarily accompanied by equivalent clinical evidence. Hua et al. (2025), in their review of 160 studies, reported that LLM-based chatbots have grown rapidly in recent literature, but only approximately 16% of studies involving LLMs had reached the stage of clinical efficacy testing. Most remained at an early validation stage. This finding highlights the gap between apparently sophisticated conversational capabilities and evidence that a system is safe and effective in therapeutic contexts. Therefore, *conversational competence* should not be treated as synonymous with *clinical competence*.

These limitations become particularly important in crisis situations. AI may recognize words associated with anxiety, depression, or hopelessness, but its responses do not necessarily account for the user's broader psychosocial context. Professional counselors, by contrast, have a responsibility to conduct risk assessments, understand contextual factors, determine the limits of their competence, and make referrals when necessary. Therefore, the more severe and complex the problems disclosed by users, the weaker the basis for positioning AI as a stand-alone source of assistance (Meadi et al., 2025; Wang et al., 2024).

The phenomenon of AI as a confidant may also alter patterns of *help-seeking*. Users may choose to speak with AI first because it is convenient, private, and less likely to evoke feelings of shame or embarrassment. This shift is not necessarily negative. AI can function as a *low-threshold entry point* that helps individuals recognize their concerns and prepare themselves before seeking human assistance. However, an important distinction must be made between AI as a bridge to help and AI as a substitute for help. In the former pattern, AI interaction may expand access to support; in the latter, comfort with AI may delay contact with counselors or mental health professionals when problems require more complex intervention.

Based on this synthesis, this article proposes the principle of a complementarity threshold, referring to a conceptual boundary for determining when AI continues to function constructively as a supportive resource and when users' needs should be transferred to human assistance. AI may be relatively appropriate for initial reflection, emotional expression, general psychoeducation, and organizing thoughts. In contrast, situations involving crisis, safety risks, complex trauma, severe psychological disorders, clinical decision-making, or the need for an ongoing therapeutic relationship should remain within the domain of professional care. Accordingly, the central issue is

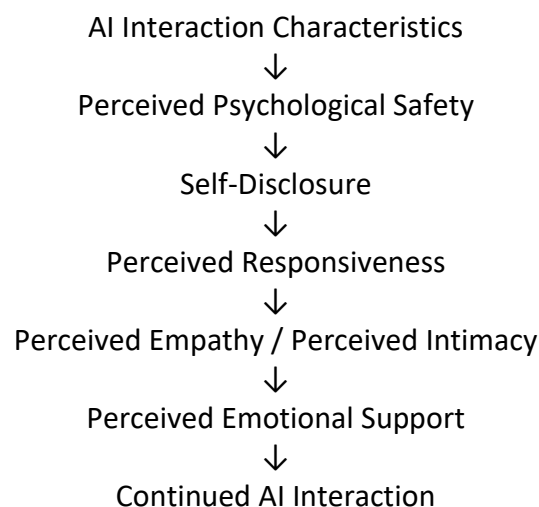
not whether AI may serve as a “confidant,” but how to ensure that closeness with AI does not obscure the boundary between digital emotional support and professional counseling.

Conceptual Framework of AI as a “Confidant” and Its Implications for Counseling

Based on the overall synthesis, the phenomenon of AI as a “confidant” can be understood as a gradual relational process rather than simply a consequence of technological sophistication. The process begins with characteristics of AI-mediated interaction—including accessibility, availability, responsiveness, reduced social evaluation, and human-like communication—which shape *perceived psychological safety*. This sense of safety reduces barriers to disclosure and encourages *self-disclosure*. When personal disclosures receive relevant, supportive, and seemingly empathic responses, *perceived responsiveness*, *perceived empathy*, and *perceived intimacy* may develop, subsequently contributing to *perceived emotional support*. This pattern helps explain how a nonhuman system can psychologically acquire a function resembling that of a “confidant” (Lee & Lee, 2023; Luo et al., 2025; Park et al., 2022).

Based on these mechanisms, this article proposes the AI-Mediated Self-Disclosure and Emotional Support Framework (AI-SDES) as a conceptual framework for explaining emotional human–AI relationships in the context of counseling. The framework does not assume that AI possesses emotions or therapeutic capacities; instead, it places users’ experiences at the center of analysis. Accordingly, emotional support is conceptualized as an experience perceived by the user, rather than as evidence that AI possesses human-like empathy.

The proposed framework can be represented as a sequence of interconnected psychological processes:



Within this framework, *anthropomorphism* functions as an important contextual factor that may strengthen, weaken, or alter the relationships among these constructs. Human-like communication may enhance *perceived responsiveness*, *perceived empathy*, and *perceived intimacy*, but excessive anthropomorphism may simultaneously increase the risk of users attributing human-like psychological capacities to AI (Lee & Lee, 2023; Zhang et al., 2025).

The framework also incorporates a complementarity threshold as a boundary condition. At lower levels of psychological complexity, AI may provide useful support for reflection, emotional expression, psychoeducation, and preparation for help-seeking. As the severity, complexity, or risk

of the user's psychological concerns increases, however, the need for professional human intervention becomes greater. Crisis situations, severe psychological distress, complex trauma, safety concerns, and clinical decision-making therefore represent contexts in which AI should not function as an independent source of counseling.

Consequently, the AI-SDES framework positions AI not as a substitute for the counselor but as a supportive layer within a broader human-centered counseling ecosystem. The conceptual contribution lies in explaining why users may experience AI as a confidant without assuming that AI possesses genuine emotional or therapeutic capacities. The framework therefore reconciles two seemingly contradictory observations: AI can generate meaningful experiences of emotional support, while simultaneously remaining fundamentally different from a professional human counselor. This distinction provides a basis for future empirical research examining the conditions under which AI-mediated self-disclosure and emotional support are beneficial, neutral, or potentially harmful.

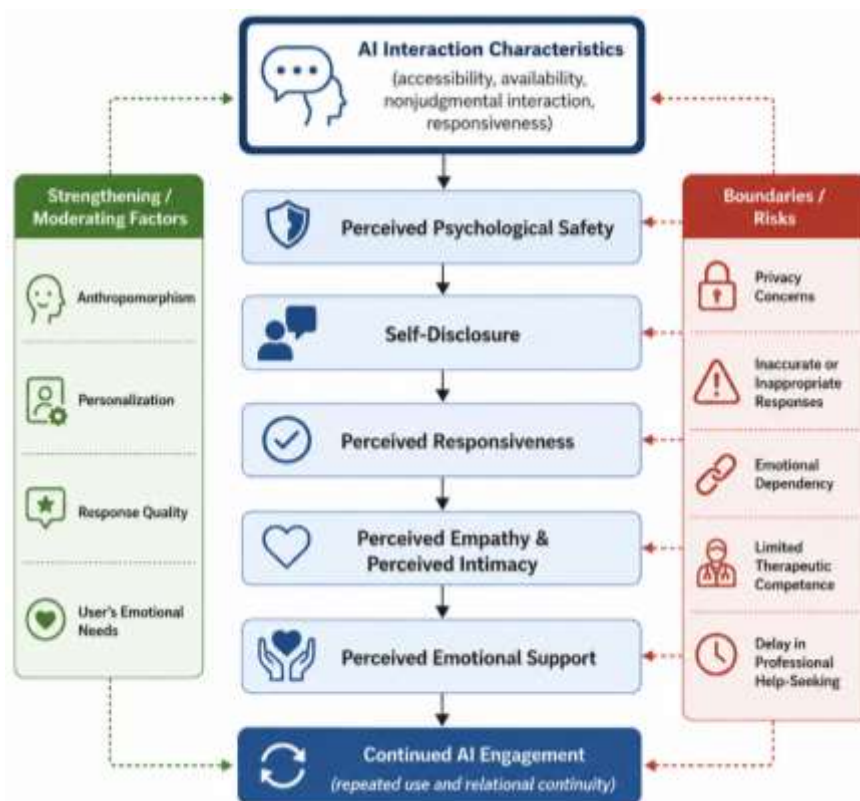


Figure 1. Conceptual Framework of AI-Mediated Self-Disclosure and Emotional Support (AI-SDES)

Source: Conceptual framework developed by the authors based on the synthesis of the literature.

The AI-SDES framework illustrates two possible pathways that may emerge after users receive emotional support from AI. The first pathway is complementary, in which AI helps users reflect on their experiences, identify emotions, obtain preliminary information, and subsequently seek interpersonal or professional support when necessary. The second pathway is substitutive, in which AI becomes the primary source of emotional support, leading users to reduce their engagement with other people or delay seeking professional assistance. This distinction is important because the benefits of AI are determined not only by how effectively a system generates responses, but also by where the interaction leads users after support has been provided.

From a counseling perspective, the complementary pathway represents the more defensible position. Meta-analytic evidence indicates that *AI-based conversational agents* have the potential to improve certain aspects of mental health, although their benefits are not consistent across all indicators of psychological well-being (Li et al., 2023). Similarly, the development of LLMs for mental health applications has progressed more rapidly than the clinical efficacy testing of these systems (Hua et al., 2025). Therefore, AI is more appropriately positioned as a *pre-counseling reflective space* and supportive technology rather than as a substitute for counselors.

This positioning has important implications for the counseling profession. Counselors need to develop AI literacy as a counseling competence. Competence in the AI era should extend beyond the ability to use technology and include the ability to explore clients' patterns of AI use, assess whether such use is adaptive or substitutive, discuss privacy risks, identify signs of emotional dependency, and help clients critically evaluate the quality of information provided by AI. Counselors should also create a safe environment in which clients can discuss their interactions with AI without fear of judgment. This approach is particularly important because the use of GenAI for emotional needs is already occurring outside formal counseling settings and is increasingly blurring the boundary between digital *self-help* and professional mental health support (Luo et al., 2025; Meadi et al., 2025).

Based on the complementarity threshold developed in this review, the position of AI within the counseling ecosystem can be operationally differentiated as presented in Table 3.

Table 3. Boundaries of AI Use within the Counseling Ecosystem

Function	Position of AI	Role of the Counselor
Initial emotional expression	Supportive	Help the client make sense of their experience
Reflection and organization of thoughts	Supportive	Deepen exploration and contextual understanding
General psychoeducation	Complementary	Verify and contextualize information
Initial screening/reflection	Limited and requires caution	Conduct professional assessment
Complex trauma/severe psychological disorders	Not a substitute	Provide professional intervention and/or referral
Crisis and safety risks	Inadequate as a sole source of support	Conduct risk assessment and provide professional intervention
Ongoing therapeutic relationship	Complementary	Counselor remains the primary therapeutic actor

Source: Authors' conceptual synthesis.

Accordingly, the primary contribution of the AI-SDES framework lies in reframing how AI should be understood within counseling. AI does not need to be positioned dichotomously as either "good" or "harmful," nor as a "new counselor" or a technology that should simply be rejected. A more appropriate perspective is to understand AI as a relational technology that can facilitate self-disclosure and generate experiences of emotional support while operating within clearly defined boundaries of competence and responsibility.

The future of counseling, therefore, is not simply a question of humans versus AI. Rather, it concerns how to develop an ecosystem of human–AI complementarity that leverages the accessibility and responsiveness of AI without compromising authentic empathy, professional judgment, ethical responsibility, and the depth of relational engagement that remain fundamental characteristics of human counseling. In this context, AI should function as an additional layer of support within the counseling ecosystem, while human counselors remain responsible for professional assessment, therapeutic decision-making, risk management, and the establishment of meaningful therapeutic relationships.

4. CONCLUSION

This conceptual review demonstrates that the use of AI as a “confidant” represents an important transformation in the human–technology relationship, in which AI is no longer used solely as an information tool but also as a medium for self-disclosure and emotional support. Through the AI-Mediated Self-Disclosure and Emotional Support (AI-SDES) framework, this review explains that AI characteristics such as accessibility, responsiveness, and perceived nonjudgmental interaction can foster psychological safety, thereby reducing barriers to self-disclosure. Self-disclosure followed by relevant and supportive responses may subsequently generate *perceived responsiveness*, *perceived empathy*, *perceived intimacy*, and ultimately *perceived emotional support*. However, feeling understood by AI should not be equated with therapeutic empathy, as AI does not possess subjective experience, ethical responsibility, or clinical competence comparable to that of professional counselors. This constitutes a central paradox of human–AI relationships: the nonhuman nature of AI may facilitate disclosure, while its human-like communication capabilities may foster emotional closeness that potentially blurs the boundary between digital support and therapeutic relationships.

In the context of counseling, AI is therefore more appropriately positioned as a *pre-counseling reflective space* and complementary technology rather than as a substitute for counselors. AI can assist individuals in expressing emotions, organizing thoughts, obtaining preliminary psychoeducation, and reducing initial barriers to help-seeking. However, its use should be guided by a complementarity threshold, particularly when concerns involve crisis situations, safety risks, complex trauma, severe psychological disorders, or the need for an ongoing therapeutic relationship. An important implication for the counseling profession is the need to strengthen AI literacy as an integral component of counselor competence, including the ability to evaluate clients’ patterns of AI use, privacy risks, information quality, and the potential for emotional dependency. Future research should empirically examine the AI-SDES framework across diverse user populations and counseling contexts to identify the conditions under which AI interaction functions as a bridge toward human support and the circumstances in which it may instead develop into a potentially harmful substitute for professional or interpersonal assistance.

DECLARATIONS

Author Contributions

Y.S: Conceptualization, literature identification and synthesis, development of the conceptual framework, analysis and interpretation, writing of the original draft, and revision of the manuscript.

M.M: Theoretical integration, critical review of the conceptual framework, interpretation of the literature, supervision, and writing, review, and editing.

N.N: Contribution to the conceptual development, literature review and interpretation, critical review of the manuscript, and writing, review, and editing.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

Ethical approval was not required because this study was a conceptual paper based exclusively on the synthesis and analysis of previously published literature and did not involve human participants, animals, or the collection of primary data.

Data Availability

No new data were generated or analyzed in this study. The study was based on previously published scholarly literature.

Acknowledgments

The authors would like to thank their respective institutions for providing academic support and access to scholarly resources that contributed to the development of this conceptual paper.

AI-Assisted Tools Disclosure

An AI-assisted tool was used solely to support the creation of the visual illustration presented in Figure 1. AI-assisted tools were not used for the conceptual development, literature analysis, interpretation of findings, or substantive writing of the manuscript. The authors reviewed and approved the final figure and remain fully responsible for the accuracy, interpretation, and content of the manuscript.

REFERENCES

- Balcombe, L. (2023, October). AI chatbots in digital mental health. In *Informatics* (Vol. 10, No. 4, p. 82). MDPI. <https://doi.org/10.3390/informatics10040082>
- Dai, X., Leng, L. L., Liu, Y., Huang, Y. T., & Wong, D. F. K. (2025). The paradox of agency in psychotherapy: How people with mental distress experience support from generative AI chatbots and human therapists. *BMC psychiatry*, 26(1), 49. <https://doi.org/10.1186/s12888-025-07671-w>
- Defossé, A., & Mullier, F. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Applied Business Research*, 104. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Haque, M. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. *JMIR mHealth and uHealth*, 11(1), e44838. <https://doi.org/10.2196/44838>
- Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., ... & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*, 24(3), 383-394. <https://doi.org/10.1002/wps.21352>
- Jaakkola, E. (2020). Designing conceptual articles: four approaches. *AMS review*, 10(1), 18-26. <https://doi.org/10.1007/s13162-020-00161-0>
- Kwesi, J., Cao, J., Manchanda, R., & Emami-Naeini, P. (2025). Exploring user security and privacy attitudes and concerns toward the use of {General-Purpose}{LLM} chatbots for mental health. In *34th USENIX security symposium (USENIX security 25)* (pp. 6007-6024).
- Lee, J., & Lee, D. (2023). User perception and self-disclosure towards an AI psychotherapy chatbot according to the anthropomorphism of its profile picture. *Telematics and Informatics*, 85, 102052. <https://doi.org/10.1016/j.tele.2023.102052>

- Li, H., Zhang, R., Lee, Y. C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1), 236. <https://doi.org/10.1038/s41746-023-00979-5>
- Luo, X., Wang, Z., Tilley, J. L., Balarajan, S., Bassey, U. A., & Cheang, C. I. (2025). Seeking emotional and mental health support from generative AI: mixed-methods study of ChatGPT user experiences. *JMIR Mental Health*, 12(1), e77951. <https://doi.org/10.2196/77951>
- Mead, M. R., Silke, T., Metselaar, S., Van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the ethical challenges of conversational AI in mental health care: scoping review. *JMIR mental health*, 12(1), e60432. <https://doi.org/10.2196/60432>
- Park, G., Chung, J., & Lee, S. (2023). Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: A serial mediation model. *Current Psychology*, 42(32), 28663-28673. <https://doi.org/10.1007/s12144-022-03932-z>
- Peng, X., Qian, Y., & He, Y. (2026). When AI chatbots understand emotions: Exploring the mechanisms of self-disclosure from emotional arousal to psychological acceptance—A hybrid SEM–ANN–NCA analysis across multiple interaction configurations. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1724313>
- Schlich, T., Barr, M., Stirman, S. W., & Raj, S. (2025). A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health*, 12(1), e69709. <https://doi.org/10.2196/69709>
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1), 46-57. <https://doi.org/10.1038/s42256-022-00593-2>
- Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., & Breazeal, C. (2024). Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study. *JMIR Mental Health*, 11(1), e62679. <https://doi.org/10.2196/62679>
- Spytska, L. (2025). The use of artificial intelligence in psychotherapy: Development of intelligent therapeutic systems. *BMC Psychology*, 13. <https://doi.org/10.1186/s40359-025-02491-9>
- Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. *Human resource development review*, 15(4), 404-428. <https://doi.org/10.1177/1534484316671606>
- Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating generative AI in mental health: systematic review of capabilities and limitations. *JMIR mental health*, 12(1), e70014. <https://doi.org/10.2196/70014>
- Zhang, Z., & Wang, J. (2024). Can AI replace psychotherapists? Exploring the future of mental health care. *Frontiers in psychiatry*, 15, 1444382. <https://doi.org/10.3389/fpsyg.2024.1444382>
- Zhang, Z., Guo, F., Fang, C., & Chen, J. (2025). Let me hold your hand: effects of anthropomorphism and touch behavior on self-disclosure intention, attachment, and cerebral activity towards AI mental health counselors. *International Journal of Human–Computer Interaction*, 41(18), 11832-11847. <https://doi.org/10.1080/10447318.2024.2446502>