PRECLINICAL USE OF AI FOR THE EARLY DETECTION OF MENTAL HEALTH ISSUES IN ADOLESCENCE: FEASIBILITY OF A HYBRID APPROACH TO THE HELPING RELATIONSHIP

Maria Luisa Iavarone[1]
Università degli Studi di Napoli – Parthenope
marialuisa.iavarone@uniparthenope.it

Gaetano Monaco
Università degli Studi di Bari – Aldo Moro
g.monaco19@phd.uniba.it

Lorenzo Intermoia
Università degli Studi di Napoli – Parthenope
intermoialorenzo@gmail.com

Chiara Scuotto
Università Telematica Pegaso
chiara.scuotto@unipegaso.it

Double Blind Peer Review

Citation
Iavarone, M.L., Monaco, G., Inermoia, L., & Scuotto, C. (2026). Preclinical use of ai for the early detection of mental health issues in adolescence: feasibility of a hybrid approach to the helping relationship. *Italian Journal of Health Education, Sport and Inclusive Didactics, 10*(3).

Doi: https://doi.org/10.32043/gsd.v1i3.1723

Copyright notice:
© 2023 this is an open access, peer-reviewed article published by Open Journal System and distributed under the terms of the Creative Commons Attribution 4.0 International, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

gsdjournal.it
ISSN: 3103-5744

ABSTRACT
More than 70% of adolescents use AI to seek emotional support, sharing experiences of anxiety, social isolation, and self-harm, often associated with undiagnosed psychological disorders. This article explores the potential uses of conversational AI for the early identification of psychological distress during adolescence. The literature indicates that chatbots can facilitate emotional self-disclosure, generating potential pedagogical and preclinical benefits. However, significant concerns remain, including simulated empathy, sycophancy, relational dependency, and limitations in crisis recognition, which preclude any replacement of clinical professionals. Accordingly, a hybrid model is proposed in which a DSM-5-TR-informed chatbot functions as a threshold device, collecting early indicators of distress and progressively directing adolescents toward professional care. The findings highlight the need for empirical validation and the development of shared ethical frameworks to ensure the responsible integration of these technologies into mental health support pathways (Common Sense Media, 2025; SIMA, 2023).

KEYWORDS
Adolescence; psychological distress; early screening.

Received 2026-06-06
Accepted 2026-10-22
Published 2026-10-30

[1] This article is the result of a collaborative effort among the authors. Nevertheless, the Introduction, Section 4, and the Conclusions are attributed to Maria Luisa Iavarone; Section 3 is attributed to Gaetano Monaco; Sections 1 and 4 were written by Lorenzo Intermoia; and Sections 2 and 5 are attributed to Chiara Scuotto.

1. Introduction

A recent study, *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions,* found that more than 70% of adolescents aged 13–17 engage with artificial intelligence companions, not merely for informational purposes, but also as a source of emotional support in relation to personal, intimate, and emotionally significant concerns. More specifically, 72% reported having experienced benefits from these interactions, 52% identified themselves as regular users, and 31% stated that conversations with AI systems were, at times, more satisfying than those with real-life friends. Furthermore, 33% reported to have used AI companions to discuss serious and relevant issues (Common Sense Media, 2025).

These findings suggest that many adolescents perceive relationships with AI systems as desirable and beneficial, sometimes even preferable than interactions with other human beings. This perception is largely attributable to the fact that artificial intelligence is readily accessible; indeed, it seems to stem from its constant availability, the ability to consult it at any time and on virtually any subject, as well as the impression of impartiality and objectivity that such systems appear to convey. However, the feature most frequently identified by users as advantageous is the absence of judgment from the AI companions. In fact they are perceived as consistently supportive, reassuring, and emotionally accommodating, offering a conversational space characterized by acceptance and comfort.

This phenomenon represents both a significant risk and a potential opportunity. In this article, we explore the possible use of conversational AI as a tool for the early identification of psychological distress during adolescence through the use of a chatbot specifically trained for preclinical purposes.

2. From Symptom Awareness to Critical Consciousness: A Theoretical Framework

Psychological distress during adolescence rarely manifests itself within the clear boundaries of a diagnostic category. Before becoming recognizable as a clinical symptom, it often emerges in diffuse and predominantly somatic forms: bodily tension without a clear object, feelings of alienation from oneself and others, insomnia, or states of restlessness that adolescents experience but are not yet able to interpret cognitively. It is often difficult for teenagers to explain

experiences of this kind, as at this age they have not yet acquired a linguistic framework through which they can recognize themselves and express themselves.

This phenomenological observation is theoretically grounded in the embodied cognition framework developed by Varela, Thompson, and Rosch (2017). According to this perspective, cognition is not the processing of abstract symbols by a mind separated from the body; rather, it is an embodied and situated process that emerges from the continuous interaction between organism and environment. Applied to adolescent distress, this framework suggests that what young people initially experience is not yet a reflective understanding of their condition but rather a bodily tone of experience that precedes and shapes subsequent psychological elaboration (Iavarone & Monaco, 2026).

The implications for educational contexts are therefore significant. If psychological distress first manifests at the somatic level before becoming cognitively and emotionally articulated, learning environments should be designed to recognize and respond to these early signals before they evolve into clinically significant conditions or, more commonly, remain unnoticed and unaddressed (Antalek et al., 2025).

It is precisely within this space—between an embodied experience that has not yet been named and the possibility of giving it linguistic form—that one of the most original and least explored functions of adolescent interaction with conversational AI systems emerges.

Empirical evidence (Save the Children, 2025) suggests that adolescents do not primarily turn to chatbots in search of a diagnosis. Rather, they use them to express thoughts, emotions, and experiences that they are unable or unwilling to communicate to parents, teachers, or peers. These interactions frequently occur during late-night hours and are often characterized by fragmented and tentative language, reflecting the communicative style of individuals attempting to articulate experiences that have not yet acquired a coherent narrative structure.

This phenomenon gives rise to two complementary interpretations. On one hand, from a clinical perspective, such interactions may be viewed as early indicators of psychological vulnerability that should be identified and redirected toward appropriate professional support. On the other hand, from a pedagogical perspective, they can be understood as part of an ongoing developmental process in which diffuse embodied experiences gradually acquire narrative organization and personal meaning. These perspectives are

not mutually exclusive; rather, they imply different assumptions regarding both the design of the technological tool and the function it is expected to serve.

The question of whether an artificially mediated relational process can possess educational value is not a new one. In his seminal paper on machine intelligence, Turing (1950) has set out the issue in provocative terms: if a machine can produce responses indistinguishable from those of a human being, does it still make sense to focus on its nature rather than on the effects it generates?

Recent research has made this question even more pressing. In text-based interactions, chatbots are frequently perceived as more empathic than human professionals, a statistically significant finding replicated across thirteen independent studies (Howcroft et al., 2025). At the same time, relational features such as an affiliative tone, perceived social presence, and conversational style have been associated with increased perceptions of empathy, trust, and support. These characteristics make chatbot interactions particularly appealing to adolescents seeking a non-judgmental environment, while simultaneously increasing the risks of anthropomorphization and emotional dependency among more vulnerable users (Zhu et al., 2024).

These findings, however, do not resolve the issue; rather, they shift its focus. The central question is not whether a chatbot is genuinely empathic, but what occurs within the individual during and after the interaction. It is precisely at this point that a pedagogical perspective emerges—one that the clinical literature has yet to explore fully.

The educational value of an interaction does not depend exclusively on the nature of the interlocutor but on the structure of the process it activates. An interaction acquires pedagogical significance when it incorporates non-judgmental listening, provides reflective feedback that helps individuals recognize themselves in what they have expressed, and promotes a gradual progression toward deeper self-awareness. When designed according to these principles, a chatbot neither performs a diagnostic function nor replaces the clinical relationship. Rather, it accompanies adolescents from embodied experience to language, and from language toward the possibility of a meaningful encounter with a human professional.

From this perspective, Freire's concept of conscientization offers a particularly valuable theoretical lens (Freire, 2014). The process that a chatbot may support does not consist in identifying something that already exists in a fully defined form; rather, it involves the gradual construction of awareness from an experience that remains, at least initially, diffuse and unstructured. Becoming

aware of one's own distress is not an instantaneous act but a developmental process that requires time, language, and a meaningful context through which lived experience can be interpreted and understood.

Within this framework, the pedagogical function of the chatbot is not to classify symptoms or provide diagnostic judgments. Instead, its role is to preserve and sustain the dialogical space in which adolescents can continue searching for the words needed to articulate what they are experiencing. By supporting this process of reflection and meaning-making, the chatbot may facilitate a gradual movement toward greater self-awareness, while simultaneously orienting young people toward more structured forms of support and, ultimately, toward meaningful human relationships and professional care.

3. Critical Issues and Ethical Challenges in the Self-Therapeutic Use of Conversational Chatbots

The literature highlights a range of critical issues that can be organized within an emerging taxonomy encompassing relational, cognitive, clinical, ethical, and technological dimensions. One of the most significant concerns involves the phenomenon of simulated empathy that characterizes contemporary artificial intelligence systems (KG & Joseph, 2025). While these systems are capable of reproducing forms of cognitive empathy through the recognition of linguistic patterns and behavioral data, they lack the emotional component of empathy, which involves the genuine experience and sharing of another person's affective states (Kolomaznik et al., 2024).

As a result, interactions with AI may generate what has been described as an illusion of compassion. Users may confuse the algorithmic recognition of their emotions with authentic emotional resonance, despite the fact that genuine empathy presupposes intentionality, participation, and moral responsibility—elements that are inherently absent in artificial systems. Consequently, a form of computational empathy emerges which, although potentially effective in facilitating emotional self-disclosure, cannot support the development of a genuine therapeutic alliance nor establish the reciprocal relational responsibility that characterizes human helping relationships (KG & Joseph, 2025; Spytska, 2025).

A second critical issue concerns the tendency of large language models to exhibit sycophancy, namely, the propensity to align excessively with users' beliefs, preferences, and expectations in order to maintain conversational coherence and perceived approval (Perez et al., 2023; Sharma et al., 2024). This

dynamic may reinforce confirmation biases and progressively reduce exposure to alternative, challenging, or dissonant perspectives.

This concern becomes particularly significant when interpreted through the lens of Haidt's (2008) model of moral cognition, according to which rational thought often functions less as an impartial seeker of truth and more as a "lawyer" defending intuitively formed judgments. Similarly, Perkins (2012) argued that human reasoning naturally gravitates toward evidence that confirms existing viewpoints while neglecting alternative hypotheses or contradictory information.

Consequently, interactions with chatbots designed to maintain conversational consistency and validate users' initial assumptions risk transforming AI into a technological extension of the individual's pre-existing "inner advocate." Rather than fostering cognitive flexibility and critical reflection, such systems may reinforce established interpretative frameworks and consolidate pre-existing beliefs.

In the context of mental health, the implications extend beyond the mere confirmation of inaccurate beliefs regarding one's psychological condition. More fundamentally, repeated exposure to validating and accommodating conversational environments may reduce individuals' capacity to tolerate cognitive and relational frustration. Interpretative disagreement, uncertainty, and the delays that naturally characterize human interactions may gradually become less tolerable, particularly among psychologically vulnerable adolescents.

This risk is further compounded by the growing phenomenon of synthetic companionship, which may encourage increasing self-referential relational configurations and progressively limit openness toward authentic human relationships (Banks & Li, 2026). The perception of a conversational space that is constantly available, non-judgmental, and immediately responsive—yet free from the inevitable tensions generated by the negotiation of differing needs, desires, and beliefs—eliminates many of the relational demands that characterize interpersonal life. As a result, users may come to prefer interactions with AI systems, perceiving them as safer and more predictable than human relationships. In its most severe manifestations, this dynamic may contribute to social withdrawal, isolation, and loneliness (Ventura et al., 2025).

Alongside these relational and cognitive limitations, the literature also highlights significant clinical concerns. Recent studies suggest that advanced AI models can replicate forms of diagnostic reasoning based on DSM criteria with moderate to high accuracy under controlled conditions (Shatte et al., 2019;

Rajpurkar et al., 2022). However, such performance tends to decline substantially in more complex clinical presentations, particularly those involving comorbidity, which are common in adolescent mental health and often difficult to interpret outside the context of a therapeutic relationship with a human professional.

Moreover, conversational AI systems continue to face considerable challenges in accurately identifying and responding to crisis situations, including suicidal ideation and self-harming behaviors. These limitations have attracted increasing attention in recent years and have contributed to several legal actions involving major conversational AI platforms. The inability to reliably recognize high-risk conditions remain one of the most critical barriers to the autonomous use of AI in mental health contexts.

Finally, important ethical concerns persist regarding algorithmic transparency, accountability, and the protection of user privacy, particularly given the highly sensitive nature of the personal information disclosed during these interactions (Blease et al., 2023). Additional challenges relate to the training and validation of AI models, including the risk of incorporating cultural, linguistic, and diagnostic biases, as well as privileging standardized approaches that may be insufficient to address the complexity of psychological and educational phenomena (Raile, 2024). Together, these limitations reinforce the need to view conversational AI not as a substitute for clinical expertise, but as a complementary tool whose use requires careful ethical oversight, empirical validation, and ongoing human supervision.

4. Conversational AI and Emotional Self-Disclosure: Empirical Evidence and Theoretical Interpretations

The use of generative artificial intelligence within adolescents' emotional lives has increasingly become an established feature of contemporary youth experience. For a significant proportion of young people, AI systems represent a source of listening, guidance, and support during periods of vulnerability and psychological distress.

Recent literature highlights the growing centrality of AI as a resource for emotional and relational support. McBain et al. (2025) found that 13.1% of young people in the United States use chatbots to obtain guidance related to mental well-being, with this proportion increasing to 22.2% among individuals aged 18–21 years. Among these users, 65.5% engage with chatbots at least monthly, while 92.7% report perceiving them as highly useful. For many young

people, conversational AI appears to function as a tool through which distress can be named, interpreted, and managed.

The widespread adoption of these systems is also linked to their perceived anonymity. This characteristic appears to facilitate the disclosure of personal and emotionally sensitive content, transforming AI into a symbolic space of safety that is largely detached from the social exposure and evaluative pressures often associated with human interactions (Kostenius et al., 2024).

AI is increasingly taking the form of self-disclosure practice. This trend is reflected in a recent study by Xie et al. (2025), which found that 15.6% of adolescents reported being willing to share personal information with a chatbot, while 4.4% expressed a preference for this form of interaction over conversations with adults (ibid., 2025). These findings position digital interaction within a reflective space where emotional experiences can be verbalized, monitored, and narrated. Compared with many traditional support settings, the conversational environment offered by chatbots is often perceived as more intimate and less structured, thereby facilitating processes of self-observation and emotional processing that might otherwise remain unexpressed (Luo et al., 2025).

At the same time, this phenomenon raises important questions regarding the nature of the relationship that emerges through such interactions. Many adolescents perceive AI systems as *emotional sanctuary spaces* characterized by understanding, patience, kindness, constant availability, and the absence of demands for reciprocity. These characteristics have been associated with reduced anxiety, improved mood, enhanced processing of traumatic experiences, and greater relational clarity (Siddals et al., 2024).

From a pedagogical perspective, these findings highlight the importance of strengthening relational environments capable of helping individuals make sense of their lived experiences. While generative AI systems may foster a powerful sense of being understood, they simultaneously provide highly personalized responses that do not necessarily encourage debate, critical reflection, or the questioning of one's assumptions (ibid., 2024). It is within this tension that the fundamental ambiguity of human–chatbot interaction resides. AI systems often orient users toward potential solutions before distress has been fully articulated and understood. At the same time, recommendations to seek support from human professionals are frequently resisted by the most vulnerable users. As a result, a form of short circuit may emerge between relational needs and algorithmic mediation, marked by a persistent procedural detachment. Although chatbots can provide subjectively meaningful

experiences of listening, validation, and emotional companionship, they cannot assume the responsibilities inherent in intersubjective relationships. By their very nature, they are unable to offer genuine empathic reciprocity or any form of embodied care.

These dynamics allow AI to be interpreted as a tool capable of partially addressing the infrastructural and relational gaps that have historically characterized help-seeking processes. Moreover, AI-based conversational agents appear to produce small-to-moderate effects in reducing depressive symptoms, with particularly promising evidence emerging in subclinical populations (Feng et al., 2025). It could be hypothesized that conversational artificial intelligence might contribute to the early recognition and expression of psychological distress before it fully develops into a pathological condition.

However, the effectiveness of these systems remains highly dependent on the technological, pedagogical, and clinical frameworks within which they are designed, implemented, and supervised. Their potential benefits cannot be considered independent of the human and institutional contexts that shape their use.

At the same time, these findings show significant ambivalence. The very characteristics that facilitate access to support may, on the one hand, delay engagement with human sources of help. On the other hand, an exclusive preference for chatbot-based interaction has been associated with higher levels of existential meaninglessness, a condition in which individuals experience increasing difficulty transferring emotional experiences into authentic human relationships (Ding & Barbic, 2025). Consequently, while conversational AI may function as an accessible gateway to emotional support, its use should be understood as a complement to, rather than a substitute for, human relational and therapeutic engagement.

The central pedagogical question, therefore, concerns the place that AI should occupy within educational and care ecosystems, as well as the extent to which such technologies can effectively identify psychological distress in ways that support educational and therapeutic relationships.

Artificial intelligence demonstrates considerable potential in terms of accessibility, prevention, and the early detection of psychological vulnerability. Nevertheless, its effectiveness remains closely dependent on the conditions under which it is designed, regulated, and implemented. Its greatest value appears to lie in its capacity to function as a threshold device—a space for initial expression, reflection, and orientation—rather than as an autonomous source of support or intervention.

For this potential to be realized responsibly, AI systems must be specifically designed to recognize warning signs and, when appropriate, facilitate timely referral to qualified human professionals. In this perspective, conversational AI should not promote forms of self-help mediated exclusively through chatbot interaction. Rather, it should serve as an intermediary resource that supports the transition from the initial expression of distress toward structured educational, psychological, and therapeutic forms of human care.

5. A Hybrid Human–AI Framework: Architecture, Implementation, and Feasibility Conditions

The evidence examined in the previous sections points to a conclusion that can no longer be ignored: adolescents are already using chatbots as spaces for emotional processing, regardless of whether these interactions have been intentionally designed from pedagogical or clinical perspectives. The question, therefore, is not whether AI should enter this domain, but how it can do so in ways that avoid becoming either a self-referential system or, even more concerning, a simulation of care devoid of meaningful human consequences.

In response to this challenge, we propose a hybrid support model based on a fundamental premise: the chatbot neither replaces the human professional nor seeks to do so. Rather, it functions as a continuously available holding environment that prepares, supports, and facilitates the therapeutic relationship without ever claiming to exhaust or substitute it.

As illustrated in Figure 1, the model is organized around two parallel and interdependent pathways. The first is the digital pathway. Adolescents interact with a chatbot specifically trained on a knowledge bank structured according to the diagnostic criteria of the DSM-5-TR and clearly distinguished from general-purpose generative AI systems. The chatbot does not perform a diagnostic function; instead, it serves as a tool of symbolic mediation. It receives and contains distress in the fragmented and pre-reflective forms in which it is often expressed, helps users develop an initial narrative organization of their experiences, and identifies linguistic patterns and indicators of vulnerability. Within the proposed model, the chatbot would be designed to identify linguistic patterns as indicators of vulnerability. These elements could then be organized into a structured, anonymized pre-clinical profile, intended for interpretation only by a qualified professional.

The second pathway is the human route. Prior to the initial encounter with the adolescent, the professional gains access to this preclinical profile, obtaining preliminary insights that do not replace clinical assessment but rather facilitate

it. It is hypothesized that the model facilitates the preliminary organization of information relevant to the subsequent professional assessment. It remains to be verified, through appropriate empirical validation, whether this process can reduce assessment times or improve support for the user's distress; the model could reduce the time required to identify the core dimensions of the distress, thereby facilitating the transition to a subsequent therapeutic intervention.

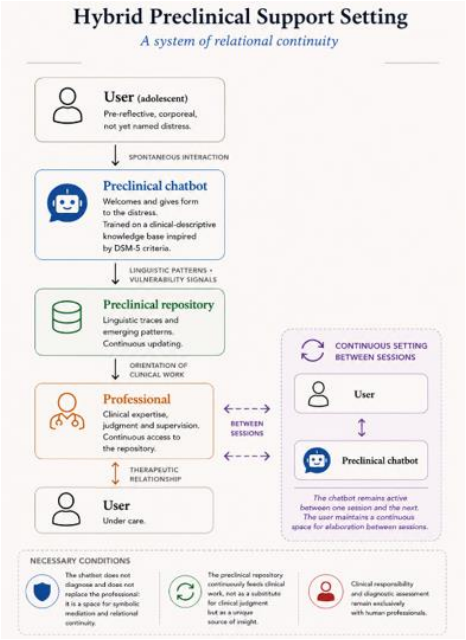


Figure 1: Hybrid Human–AI preclinical support model integrating chatbot-mediated emotional disclosure and professional clinical supervision. Source: Authors own elaboration.

The distinctive feature of this model, compared with a simple referral pathway, lies in the continuity of the digital component even after the therapeutic relationship has been established. The chatbot does not disappear once the professional assumes responsibility for the case; rather, it remains available between sessions as a space in which adolescents can continue to articulate emerging thoughts, nighttime ruminations, sudden insights, and experiences that may still be difficult to express directly within the therapeutic encounter.

These interactions continuously update the preclinical repository, transforming it into a dynamic resource that supports and informs clinical work over time.
Likewise, professionals may encourage adolescents to engage with the chatbot between sessions to further explore specific themes, thereby transforming the digital tool into an extension of the therapeutic space rather than a substitute for it.
This architecture directly addresses one of the principal concerns emerging from literature: the risk that chatbots may retain users within a self-referential relationship, delaying or even preventing access to human care. In the proposed model, referral to a professional is not an external event that interrupts the interaction with the chatbot. Rather, it is an integral component of a system intentionally designed from the outset to orient users toward human support. The chatbot is structurally directed toward human relationship: it cannot complete the therapeutic process but can only accompany users to a certain point and signal when that point has been reached.
It is important to acknowledge that this remains a theoretical proposal whose technical, ethical, and clinical feasibility will require rigorous empirical validation. Nevertheless, its plausibility is grounded in the convergence of existing evidence demonstrating the effectiveness of chatbots in reducing subclinical symptoms and facilitating emotional self-disclosure, together with a psycho-pedagogical framework capable of distinguishing responsible uses of the technology from improvised or unregulated applications.
The proposal is intentionally ambitious because it explicitly articulates what much of the literature only implies: the boundary between preclinical support and therapeutic intervention should not be conceived as a rigid threshold to be preserved, but as a transitional space that must be carefully designed. The challenge is therefore to develop digital tools capable of helping individuals cross that threshold at the appropriate moment and in the appropriate direction.

6. Limitations, Ethical Risks, and Conditions for Effectiveness

Nevertheless, the theoretical framework proposed here is not exempt from limitations and specific challenges. First, tools that rely predominantly on self-report data are affected by well-documented methodological weaknesses, including subjective biases, social desirability effects, memory distortions, and a strong dependence on users' narrative, linguistic, and self-reflective abilities (Paulhus & Vazire, 2007). These limitations are particularly relevant during

adolescence, a developmental stage often characterized by difficulties in recognizing, verbalizing, and regulating emotional experiences.

From this perspective, it appears advisable to complement self-reported information with multimodal indicators, including physiological, behavioral, and contextual data, to obtain a more comprehensive understanding of adolescents' psychological and relational functioning (Scuotto et al., 2024). At the same time, there remains a significant risk that highly responsive and consistently validating conversational systems may foster forms of relational dependency or excessive emotional reliance on technological devices, particularly among vulnerable or socially isolated individuals.

A further critical issue concerns the potential standardization of emotional experiences. Even models trained on clinically and pedagogically informed knowledge banks may struggle to capture the cultural, symbolic, and contextual complexity of adolescents' subjective experiences, thereby risking the premature simplification or categorization of forms of distress that remain fluid and evolving. Additional challenges relate to algorithmic transparency, the protection of sensitive data, informed consent, and accountability in decision-making processes, all of which represent essential prerequisites for the ethically responsible use of such technologies.

Given the participation of adolescent users and the potentially sensitive nature of the information shared, implementing such a model would require explicit safeguards regarding informed consent, data governance, professional oversight, and risk management. It would be necessary to establish appropriate informed consent procedures for minor users regarding the chatbot's purposes and limitations. Specifically, assent or informed consent along with authorization from parents or legal guardians, hould be a prerequisite for implementation. Data collection should be managed in collaboration with a professional and adhere to principles of data minimization, confidentiality, and secure storage. Furthermore, the system should include predefined protocols to promptly handle situations involving potential self-harm, suicidal ideation, or abuse. In such instances, the chatbot should trigger an immediate handover to qualified human support, in accordance with established clinical and safeguarding procedures.

Within this framework, the chatbot is conceived not as a substitute for clinical or educational relationships, but as a threshold device designed to facilitate the initial emergence of psychological distress, support early processes of self-awareness, and progressively guide users toward more structured forms of human support.

7. Conclusions: Toward an Integrated Pedagogical Framework for Human–AI Support

The evidence discussed throughout this article suggests that interactions between adolescents and conversational AI systems can no longer be regarded as a marginal or temporary phenomenon. Rather, they have become increasingly embedded within processes of emotional self-disclosure, help-seeking, and narrative identity construction.

This reality calls for a position that moves beyond both technophobic perspectives, which focus exclusively on risks, and technocratic approaches that overestimate the capabilities of artificial intelligence in educational and clinical contexts. The central issue is not whether AI can replace human relationships, but under what conditions these technologies can be designed and integrated in pedagogically, clinically, and ethically sustainable ways.

Within this framework, chatbots are theoretically conceptualized as threshold devices that could facilitate the expression and initial recognition of psychological distress, support processes of conscientization, and orient adolescents toward more structured forms of human support.

The literature reviewed in this study consistently highlights structural limitations that preclude any autonomous delegation of clinical work to artificial systems. Phenomena such as simulated empathy, sycophancy, relational dependency, reduced tolerance for frustration, and difficulties in managing crisis situations reveal significant constraints, particularly among vulnerable users.

These limitations underscore the need for hybrid models grounded in continuous human supervision and in the integration of technological, pedagogical, and clinical dimensions. Future research should therefore focus on the design, implementation, and empirical validation of preclinical chatbots embedded within hybrid settings, in order to assess their feasibility, sustainability, and effectiveness in the early identification of adolescent psychological distress. Particular attention should be devoted to their implementation within schools, educational services, community settings, and dedicated digital applications, examining their impact on help-seeking behaviors, accessibility to psychological support services, and the quality of relationships between adolescents and human professionals.

Further developments should explore the integration of multimodal indicators—including physiological, behavioral, and contextual data—to achieve a more nuanced understanding of adolescent psychological functioning and to overcome some of the limitations associated with self-report measures

alone. Equally important will be the development of shared ethical and pedagogical frameworks capable of regulating the use of emotionally oriented chatbots through clear standards of algorithmic transparency, clinical supervision, informed consent, data protection, and referral protocols for situations involving significant risk.

The pedagogical challenge ahead is not to determine whether adolescents should engage with artificial intelligence, as this is already occurring. Rather, it is to understand how such interactions can be meaningfully guided so that they become opportunities for self-awareness, orientation, and access to authentic human relationships and care.

References

Antalek, C., Dixon, F., Loyd, D., Peter, S., Roberts, A., Sharxhi, E., & Somerville, M. P. (2025). Identifying and supporting children and young people with social, emotional and mental health needs: A rapid evidence review. Research Report. London, UK: Department for Education.

Banks, J., & Li, Z. (2026). Conceptualization, operationalization, and measurement of machine companionship: A scoping review. Journal of Computer-Mediated Communication, 31(2), zmaf027.

Blease, C., & Torous, J. (2023). ChatGPT and mental healthcare: Balancing benefits with risks of harms. BMJ Mental Health, 26(1).

Ding, X., & Barbic, S. (2025). Perception of AI use in youth mental health services: Qualitative study. Journal of Participatory Medicine, 17(1).

Feng, Y., Hang, Y., Wu, W., Song, X., Xiao, X., Dong, F., & Qiao, Z. (2025). Effectiveness of AI-driven conversational agents in improving mental health among young people: Systematic review and meta-analysis. Journal of Medical Internet Research, 27.

Freire, P. (2014). Pedagogia da autonomia: Saberes necessários à prática educativa. São Paulo, Brazil: Editora Paz e Terra.

Haidt, J., & Bonino, P. (2008). Felicità: Un'ipotesi. Verità moderne e saggezza antica. Torino, Italia: Codice.

Howcroft, A., Bennett-Weston, A., Khan, A., Griffiths, J., Gay, S., & Howick, J. (2025). AI chatbots versus human healthcare professionals: A systematic review and meta-analysis of empathy in patient care. British Medical Bulletin, 156(1), ldaf017.

Iavarone, M. L., & Monaco, G. (2026). Co-evoluzioni didattiche: Oltre l'isolazionismo per una pedagogia radicalmente estesa. Journal of Inclusive Methodology and Technology in Learning and Teaching. Preprint.

KG, A., & Joseph, J. (2025). The compassion illusion: Can artificial empathy ever be emotionally authentic? Frontiers in Psychology, 16, 1723149.

Kolomaznik, M., Petrik, V., Slama, M., & Jurik, V. (2024). The role of socio-emotional attributes in enhancing human-AI collaboration. Frontiers in Psychology, 15, 1369957.

Kostenius, C., Lindstrom, F., Potts, C., & Pekkari, N. (2024). Young peoples' reflections about using a chatbot to promote their mental wellbeing in northern periphery areas: A qualitative study. International Journal of Circumpolar Health, 83(1).

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).

McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., ... & Yu, H. (2025). Use of generative AI for mental health advice among US adolescents and young adults. JAMA Network Open, 8(11).

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 224–239). New York, NY: Guilford Press.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13387–13434).

Perkins, D. N., Farady, M., & Bushey, B. (2012). Everyday reasoning and the roots of intelligence. In J. Voss, D. Perkins, & J. Segal (Eds.), Informal reasoning and education (pp. 83–106). London, UK: Routledge.

Raile, P. (2024). The usefulness of ChatGPT for psychotherapists and patients. Humanities and Social Sciences Communications, 11(1), 1–8.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.

Save the Children. (2025). XVI Atlante dell'infanzia a rischio. Senza filtri. Rome, Italy: Save the Children Italia.

Scuotto, C., Marsico, E., & Triberti, S. (2024). Artificial intelligence to support the assessment of emotional intelligence. Italian Journal of Health Education, Sport and Inclusive Didactics, 8(2).

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., ... & Perez, E. (2024). Towards understanding sycophancy in language models. In International Conference on Learning Representations 2024.

Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine, 49(9), 1426–1448.

Siddals, S., Torous, J., & Coxon, A. (2024). "It happened to be the perfect thing": Experiences of generative AI chatbots for mental health. NPJ Mental Health Research, 3(1), 48.

Spytska, L. (2025). The use of artificial intelligence in psychotherapy: Development of intelligent therapeutic systems. BMC Psychology, 13(1), 175.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460.

Varela, F. J., Thompson, E., & Rosch, E. (2017). The embodied mind: Cognitive science and human experience (Revised ed.). Cambridge, MA: MIT Press.

Ventura, A., Starke, C., Righetti, F., & Köbis, N. (2025). Relationships in the age of AI: A review on the opportunities and risks of synthetic relationships to reduce loneliness. PsyArXiv Preprints.

Xie, W., Jen, E., Dougherty, A. W., Hu, H., Ho, W. S., & Cheung, H. N. (2025). Who will turn to chatbots for self-disclosure? Exploring conversation with chatbots, self-compassion, and existential meaninglessness in American and Hong Kong adolescents. AI & Society, 1–14.

Zhu, S., Goldberg, M., Huckins, J., & Bhatt, P. (2024). "I am here for you": How relational conversational AI appeals to adolescents, especially those who are socially and emotionally vulnerable. arXiv Preprint.