Adebayo, B., & Wang, S. (2026). AI as emotional support in higher education: Opportunities, ethical risks, and institutional responsibility. *International Journal of Technology in Teaching and Learning, 22(1)*, 51-68.

AI as Emotional Support in Higher Education: Opportunities, Ethical Risks, and Institutional Responsibility

Babatunde Adebayo
Shuyan Wang
University of Southern Mississippi

Universities are adopting artificial intelligence (AI) conversational agents that offer students emotional support faster than they are building the rules needed to govern them. This position paper argues that the features that make these systems valuable are the same features that make them risky. Constant availability, anonymity, human-like warmth, and data-driven scale each offer potential benefits while also creating corresponding governance concerns. Availability widens access but may reduce opportunities or prompts for human help-seeking; anonymity lowers the barrier to disclosure but can separate disclosure from a person who can respond; warmth sustains engagement, but invites misplaced trust and dependency; and scale extends reach at low cost but concentrates privacy exposure and can amplify model limitations across student populations. Because benefit and risk originate in the same property, technical safeguards alone are unlikely to address the institutional responsibilities associated with these systems and must therefore be complemented by governance. Drawing on randomized trials, meta-analytic and review evidence, campus deployments, and established ethical and psychological theory, the paper shows that evidence of benefit is concentrated among students with mild to moderate distress over short periods. In contrast, evidence on dependency, crisis response, long-term outcomes, and equitable performance across student groups remains limited. The paper, therefore, advances a position in which universities adopt AI emotional support only as a bounded supplement to human care, within a proposed tiered governance framework that specifies triggering criteria, required actions, responsible roles, and escalation routes, and that keeps human professionals accountable for decisions involving clinical judgment, safety, escalation, or significant student well-being. The paper responds to five counterarguments and concludes with recommendations for policy, practice, and research.

Babatunde Adebayo is a PhD student and Dr. Shuyan Wang is a Professor from the School of Leadership at the University of Southern Mississippi. Babatunde Adebayo can be reached at babatunde.adebayo@usm.edu
DOI: https://doi.org/10.37120/ijttl.2026.22.1.04

Keywords: Artificial intelligence, emotional support, higher education, student mental health, institutional governance, stepped care.

INTRODUCTION

Student mental health needs in higher education now exceed what campus counseling services were built to provide. Data from more than 350,000 students at 373 campuses in the United States show that in 2020–2021, over 60% met the criteria for at least one mental health challenge, an increase of nearly 50% since 2013, and that gaps in treatment persisted for students of color (Lipson et al., 2022). The consequences reach beyond well-being; depression predicts a lower grade point average and a higher probability of dropping out, particularly when it co-occurs with anxiety (Eisenberg et al., 2009). Into this gap, universities and students have brought a new tool. Artificial intelligence (AI), meaning digital systems that perform tasks associated with human cognition, such as language processing and learning from data (Russell & Norvig, 2021), is no longer confined to tutoring, assessment, and administration in higher education (Ouyang et al., 2022; Zawacki-Richter et al., 2019). Conversational agents now offer encouragement, coping strategies, and a sympathetic reply at any hour, and research on AI in education has grown rapidly over the past decade (Guo et al., 2024).

The challenge now is that adoption is outpacing governance. A scoping review of 118 studies on large language models in education identified recurring weaknesses in transparency, replicability, and attention to privacy (Yan et al., 2024), and UNESCO guidance reports that most educational institutions are not yet equipped to validate generative AI tools before using them (Miao & Holmes, 2023). An earlier systematic review of AI in higher education found little critical reflection on ethical risk within the research base itself (Zawacki-Richter et al., 2019). This gap matters most where the data are most sensitive, and few data held by a university are more sensitive than a student's private account of their own distress.

The central claim of this paper is that the properties that make AI emotional support attractive to universities, namely constant availability, anonymity, human-like warmth, and data-driven scale, are also the properties that can generate distinctive governance challenges. Availability broadens access but may reduce opportunities for human contact when students rely repeatedly on an automated system. Anonymity lowers the barrier to disclosure but can separate disclosure from the person who can interpret and respond to it. Warmth sustains engagement, but invites misplaced trust and dependency. Scale extends reach at low cost but can magnify privacy risks and reproduce model limitations across the institution. Because each benefit and its corresponding risk arise from the same design property, technical design can mitigate these risks but cannot, by itself, determine the institutional boundaries, responsibilities, and escalation procedures governing their use.

The position that follows is therefore specific: universities should adopt AI emotional support only as a bounded supplement to human care, within a tiered model that keeps human professionals accountable for decisions involving clinical judgment, safety, escalation, or significant student well-being. This position applies primarily to AI systems that universities provide, license, embed, or formally endorse for student support; it does not attempt to govern every general-purpose AI system that students may independently choose to use. The existing literature supports part of this claim but leaves another part open. What is established is that structured conversational agents can reduce depressive symptoms and psychological distress in the short term, particularly among young adults with mild to moderate symptoms (Fitzpatrick et al., 2017; Li et al., 2023). The evidence is mixed regarding anxiety and broader well-being, and it is thin on long-term effects, dependency, and how these systems perform with students at elevated risk (Laestadius et al., 2024; Li et

al., 2023; McBain et al., 2025). What is missing is an account that connects benefit and risk rather than listing them separately, and that translates this relationship into a governance framework that institutions can evaluate and implement. Ethical frameworks for AI are well developed at the level of principle (Floridi & Cowls, 2019; Holmes et al., 2022), but they do not specify in sufficient operational detail who acts, on what signal, and through which route when a student's conversation with a machine becomes a matter of concern.

This paper addresses that gap by organizing the evidence around the design properties so that each benefit is read alongside the risk that comes with it (Table 1). The paper consistently distinguishes among empirical findings, theoretical mechanisms, and normative recommendations. It then tests the resulting position against counterarguments and translates the position into a proposed tiered governance framework with explicit triggers, actions, responsibilities, and escalation routes (Table 2). This is a position paper rather than an empirical study, and its argument rests on published trials, systematic reviews, campus deployments, and established ethical and psychological theory. The campus deployments are used primarily as evidence about bounded institutional uses of conversational AI rather than as direct evidence of emotional-support outcomes. The paper moves from the literature review, which sets out what is known and what is missing, to the central argument, the counterarguments, the recommendations, and the limitations and directions for future research. Figure 1 summarizes the logic of the argument.

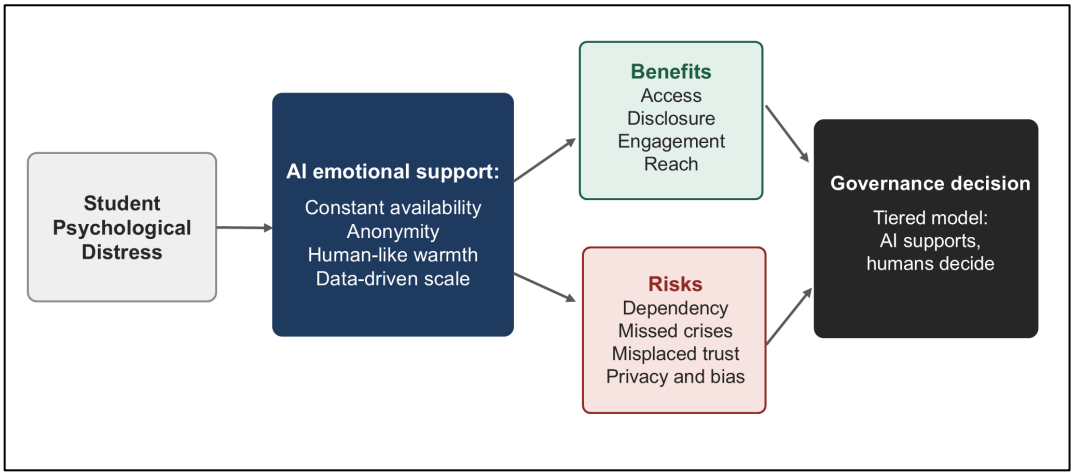


Figure 1. Designed Properties for AI Emotional Support Benefits and Risks

LITERATURE REVIEW

This review brings together four bodies of work: randomized trials and meta-analyses on conversational agents for mental health; evidence from campus deployments; theory from psychology and human-computer interaction; and scholarship on the ethics and governance of AI in education. It is organized to move from what is settled to what is not. It first defines AI emotional support and its main forms in higher education, then sets out evidence of benefit, areas where that evidence is mixed or limited, theories explaining why benefit and risk are linked, and the current state of governance guidance. It closes by identifying the specific gaps that the remainder of this paper addresses. Throughout this review, evidence from therapeutic or mental-health-specific systems is distinguished from evidence involving general educational chatbots or general-purpose companion systems, because findings from one category cannot automatically be generalized to another.

DEFINING AI EMOTIONAL SUPPORT IN HIGHER EDUCATION

In this paper, AI emotional support refers to conversational systems that recognize or respond to students' affective states by offering reassurance, coping strategies, encouragement, or empathic dialogue (Wollny et al., 2021). Three forms are present in higher education. The first is the purpose-built therapeutic agent, which delivers structured techniques drawn from cognitive behavioral therapy, as in the trials reviewed below. The second is the embedded agent, which sits inside coursework or student services and provides socioemotional scaffolding as part of another task, such as goal setting in an online course (Hew et al., 2023). The third is the general-purpose or companion chatbot, which students use on their own initiative for emotional conversation even though it was not designed or evaluated for that purpose (Laestadius et al., 2024). In their systematic review of chatbots in education, Wollny et al. (2021) found mentoring roles, including motivational support, alongside the more common learning and assisting roles.

The distinction among these forms matters for governance because their purposes, evidence bases, and institutional responsibilities differ. Evidence from a purpose-built therapeutic agent should not be treated as evidence that a general-purpose companion chatbot is equally safe or effective for emotional support. Similarly, evidence that an embedded educational chatbot can provide motivational or social support does not establish that it can perform a mental-health assessment or intervention. The position advanced in this paper, therefore, concerns the governance of these systems according to their intended use and the level of emotional or clinical responsibility that their use creates.

BENEFITS OF AI EMOTIONAL SUPPORT

The documented benefits of AI emotional support are most clearly established for short-term symptom reduction in structured interventions, rather than for broad claims about student well-being, retention, or long-term mental health. In a randomized trial with 70 young adults, a fully automated agent delivering cognitive behavioral techniques over two weeks produced a significantly greater reduction in depressive symptoms than an information-only control, with a moderate between-groups effect (Fitzpatrick et al., 2017). A trial that recruited students from 15 universities across the United States reported reduced symptoms of depression and anxiety relative to a control condition (Fulmer et al., 2018). In a 16-week trial with 83 university students, a chatbot-delivered self-help intervention reduced depressive symptoms more than bibliotherapy did (Liu et al., 2022). At the level of synthesis, a meta-analysis of 15 randomized trials found that AI-based conversational agents significantly reduced symptoms of depression and psychological distress, with larger effects for generative than for retrieval-based agents (Li et al., 2023).

These findings support the narrower claim that some structured conversational interventions can provide useful low-intensity support under defined conditions. They do not, by themselves, establish that open-ended emotional companionship, general-purpose chatbots, or institution-wide deployment produces equivalent benefits. This distinction is important because the proposed position does not reject the evidence of benefit; rather, it uses the limits of that evidence to define the boundaries within which institutional adoption can be justified.

Evidence from campus practice provides additional insight into the value of bounded conversational AI. However, these studies should be interpreted as indirect evidence for the governance argument rather than as evidence of emotional or clinical benefit. Georgia State University introduced a text-based conversational assistant to help admitted students complete the tasks standing between acceptance and the first day of class. In a field experiment, students assigned to the assistant completed pre-enrollment requirements more successfully. They were 3.3 percentage points more likely to enroll on time, with a

substantially lower burden on university staff than comparable human outreach would have required (Page & Gehlbach, 2017). The system handled routine, high-volume questions; its purpose was clearly bounded, and matters beyond its scope were left to human staff. In instruction, two studies in fully online courses used chatbots to guide students through goal setting and serve as learning companions, with a focus on social presence. Students reported positive experiences and substantial engagement (Hew et al., 2023).

The relevance of these campus examples is therefore not that they demonstrate the effectiveness of AI emotional support. Rather, they illustrate a recurring institutional principle: conversational AI can be useful when its purpose is clearly bounded, its limitations are recognized, and human support remains available for needs outside its scope. That principle informs the governance argument developed later in the paper.

LIMITATIONS OF THE BENEFITS OF AI EMOTIONAL SUPPORT

The evidence for benefit is promising but bounded. In the Woebot trial, anxiety declined in both the intervention and the control conditions (Fitzpatrick et al., 2017). In the meta-analysis by Li et al. (2023), effects on anxiety and overall psychological well-being were not statistically significant; effects in non-clinical populations were small; and only six studies reported follow-up effects. The authors graded the evidence on distress as of moderate quality due to substantial heterogeneity. Most included studies were single-site and had fewer than 200 participants. Evidence that AI emotional support improves retention or degree completion, as distinct from symptom reduction, is limited and should not be inferred from symptom reduction alone.

Student perspectives are also divided. In a survey of university students in Hong Kong, respondents valued generative AI for its accessibility and personalized support. However, they raised concerns about its accuracy, privacy, ethics, and effect on their personal development (Chan & Hu, 2023). Cultural context shapes these views. In many Asian higher education settings, stigma surrounding mental health and the social importance of preserving face may influence whether and how students disclose distress, potentially making anonymous AI support attractive to some students while raising distinct questions about privacy, interpretation, and responsibility (Day, 2023). These cultural considerations should be treated as context-sensitive influences rather than generalized assumptions about Asian students as a group.

The evidence on the challenges of AI as an emotional companion remains limited, but the available literature identifies several concerns that warrant precaution. A grounded theory analysis of 582 posts by users of the companion chatbot Replika documented mental health harms arising from emotional dependence, including role-taking, in which users came to feel that the chatbot had needs of its own requiring their attention (Laestadius et al., 2024). Because this study examined posts from users of a commercial companion chatbot rather than students using a university-provided intervention, its findings cannot be directly generalized to higher education. It does, however, provide evidence that emotionally oriented human–AI relationships can develop in ways that raise questions about dependency and boundaries.

An evaluation of three widely used large language model chatbots, which submitted 30 suicide-related queries 100 times each to each system, found that the chatbots responded in line with clinician judgment at the very-low and very-high ends of risk but showed limitations in distinguishing intermediate levels of risk (McBain et al., 2025). This finding is particularly relevant to governance because risk identification in real conversations is unlikely to consist only of clearly low- or clearly high-risk statements. However, the evaluation should be interpreted as a controlled test of chatbot responses rather than as evidence of how students would experience or use these systems in practice.

Of the 35 studies in the review by Li et al. (2023), fewer than half incorporated any safety assessment or protective measure. Meanwhile, randomized trials of therapeutic agents generally exclude participants at elevated risk, for sound ethical reasons. The result is an important limitation on the evidence: the literature provides more information about short-term outcomes among participants with relatively manageable levels of distress than about performance in the higher-risk situations in which institutional accountability is most consequential. This limitation does not establish that AI systems will fail in such situations; rather, it means that institutions should not assume that evidence from lower-risk populations establishes safety for higher-risk use.

THEORETICAL FOUNDATIONS: SOCIAL RESPONSE AND SELF-DETERMINATION THEORIES

Two bodies of theory explain why benefit and risk are linked. Social response theory holds that people apply social rules to computers that display human-like cues, often mindlessly and even when they know they are interacting with software (Nass & Moon, 2000). Later work confirmed that human-like design cues in chatbots increase users' compliance with the system's requests (Adam et al., 2021). These findings provide a theoretical mechanism through which human-like design may increase engagement while also increasing the possibility that users place more trust in a system than its actual competence warrants. They do not establish that every warm or anthropomorphic chatbot produces dependency or inappropriate trust.

Self-determination theory identifies autonomy, competence, and relatedness as basic psychological needs whose satisfaction supports motivation and well-being (Ryan & Deci, 2017). The theory provides an interpretive lens rather than an empirical test of the AI systems reviewed here. For any tool, the question is whether a student who uses it over a full semester becomes more able to manage distress independently, more confident in their own competence, and more connected to other people, or less so on each count. From this perspective, a system that supports these capacities may complement human care. In contrast, a system that consistently substitutes for opportunities to develop them could undermine the conditions associated with well-being. This distinction remains a proposition for empirical investigation rather than an established outcome of AI emotional support.

Together, these theories explain the central tension developed in this paper. Social response theory helps explain why human-like warmth can increase engagement and trust; self-determination theory provides a framework for asking whether that engagement ultimately supports or displaces students' autonomy, competence, and human relatedness. Neither theory, however, establishes the effectiveness or safety of a particular AI system. Their role in this paper is to explain plausible mechanisms and identify outcomes that governance and future research should examine.

ETHICAL AND GOVERNANCE FRAMEWORK

The governance literature is more developed at the level of principle than at the level of practice. Floridi and Cowls (2019) synthesized prominent statements of AI ethics into five principles: beneficence, non-maleficence, autonomy, justice, and explicability, the last of which combines intelligibility with accountability. In education, Holmes et al. (2022) distinguished between doing ethical things and doing things ethically, arguing that developers' good intentions are insufficient and that fairness, accountability, transparency, bias, autonomy, agency, and inclusion must each be addressed explicitly. UNESCO guidance proposes a human-centered approach in which the protection of users' data privacy is mandated, and generative AI tools are ethically and pedagogically validated before use (Miao & Holmes, 2023).

Empirical reviews show how far practice may lag behind these principles. Research on large language models in education pays insufficient attention to privacy, transparency, and beneficence (Yan et al., 2024), and research on AI in higher education has engaged little with ethical risk (Zawacki-Richter et al., 2019). These findings support the need for stronger governance attention, but they do not by themselves demonstrate that a particular institution has failed to govern a particular emotional-support system.

What these sources share is a consensus on values. What they do not provide in sufficient operational detail for the specific problem addressed here is a mechanism that links a particular design property to a particular harm and then identifies the human role responsible for monitoring, mitigating, or responding to that risk. The contribution proposed by this paper is therefore not another list of AI ethics principles. It is an attempt to translate established principles into an institutional decision structure for AI emotional support.

THE GAP IN LITERATURE

The gaps identified in the review are conceptual, empirical, and operational. First, the literature often presents benefits and risks as separate categories (e.g., Chan & Hu, 2023). Such a structure can imply that benefits and risks are independent characteristics that can be balanced against one another. The present paper instead proposes that at least some of the relevant benefits and risks arise from the same design properties. Constant availability can increase access while also creating opportunities for repeated reliance; anonymity can facilitate disclosure while separating disclosure from a human responder; warmth can increase engagement while potentially increasing misplaced trust; and scale can extend access while magnifying privacy and equity concerns.

Second, the empirical evidence remains limited in several areas relevant to institutional adoption. Evidence is stronger for short-term symptom reduction than for long-term effects, dependency, crisis response, or equitable performance across student populations (Li et al., 2023; McBain et al., 2025). The appropriate conclusion is not that these systems are unsafe in all such circumstances, but that the current evidence does not justify assuming safety or effectiveness where direct evidence is sparse.

Third, the literature is operationally incomplete. Existing frameworks establish values that should guide AI adoption but provide less detail about how an institution should translate those values into deployment boundaries, escalation procedures, assigned responsibilities, and evaluation criteria (Holmes et al., 2022; Miao & Holmes, 2023). A counseling center, student affairs office, faculty member, or procurement team may understand that privacy, autonomy, fairness, and safety matter without having a clear framework for determining who is responsible for each issue when a specific system is deployed.

This paper addresses these gaps in three corresponding ways. It addresses the conceptual gap by analyzing benefits and risks as consequences that can arise from shared design properties; the empirical gap by restricting the proposed institutional role of AI to uses that are reasonably aligned with the existing evidence and by identifying areas requiring further study; and the operational gap by proposing a tiered governance framework that assigns triggers, actions, responsibilities, and escalation routes. The framework is presented as a conceptual proposal derived from the literature, not as an empirically validated model.

THE POSITION: DESIGN PROPERTIES AND CONSEQUENCES

This section develops the central claim property by property. Table 1 pairs each design property with the benefits and potential risks it produces, the evidence supporting each, and the governance response it implies. The subsections that follow explain each row, and the final subsection explains why design alone cannot resolve the tension that the table displays. The argument applies to all three forms of AI emotional support defined earlier, but with

different force. In purpose-built therapeutic agents, availability and warmth are deliberate design choices that can be directly governed. In embedded agents, the task boundary already limits the emotional stakes. In general-purpose companions, the institution controls neither the design nor the data, so its governance acts through what it endorses, how it signposts human services, and how it prepares students to judge these tools for themselves. The relationships summarized in Table 1 should therefore be understood as potential benefits, risks, and governance implications rather than outcomes that necessarily occur in every system or for every student.

Table 1. *Design Properties of AI Emotional Support*

Design Property	Benefit	Potential Risk or Concern	Proposed Governance Response
Constant availability	Immediate, low-barrier support without waitlists or scheduling (Fitzpatrick et al., 2017; Fulmer et al., 2018; Liu et al., 2022)	Fewer natural prompts toward human support; potential emotional dependence (Laestadius et al., 2024)	Prompts toward human contact in extended or repeated sessions; Tier 1 limited to evidence-supported uses; review of aggregate usage patterns
Anonymity	Lower stigma and an easier first disclosure (Day, 2023)	Disclosure may occur without a person who can act on it; informed consent may be difficult to secure; intermediate-risk disclosures may be missed (McBain et al., 2025)	Crisis exit is designed and tested before launch; a wide escalation threshold; a route to a person that functions whether or not the student is identified.
Human-like warmth	Engagement and a perceived therapeutic relationship (Hew et al., 2023; Li et al., 2023)	Trust that may exceed demonstrated competence; compliance; potential dependence (Adam et al., 2021; Laestadius et al., 2024; Nass & Moon, 2000)	Recurring disclosure of AI status and limits; no design features that simulate the system's own needs
Data-driven scale	Reach large and underserved populations at low marginal cost (Lipson et al., 2022; Page & Gehlbach, 2017)	Cumulative privacy exposure; model bias or unequal performance may be amplified across the institution (Day, 2023; Dwork & Roth, 2014; Yan et al., 2024)	Retention and secondary-use limits in vendor contracts; right to audit; outcomes disaggregated by student group

The table presents potential relationships rather than deterministic outcomes. Some are supported directly by empirical evidence, whereas others are informed by theory or by evidence from adjacent forms of conversational AI. This distinction is particularly important because long-term evidence specific to university-provided AI emotional support remains limited.

CONSTANT AVAILABILITY

Availability is the strongest argument for AI emotional support. A student awake with worry at 2 a.m., or waiting weeks for a counseling appointment, can receive structured coping support immediately, and the trials reviewed above show that brief, structured support of this kind can reduce depressive symptoms (Fitzpatrick et al., 2017; Fulmer et al., 2018). However, availability may also reduce some of the friction that ordinarily turns a student toward other people. A counseling session ends, an office closes, and a friend goes to sleep; each such moment is a time when a student may reach out to someone else. A system that never closes offers no such natural stopping point.

The documented dependency among companion chatbot users raises potential concerns, as some users have come to organize their emotional lives around the system and feel responsible for its needs (Laestadius et al., 2024). Because this evidence comes from a commercial companion chatbot rather than a university-provided intervention, it should be treated as indirect evidence of what sustained reliance can look like rather than as evidence that such dependency will occur in higher education. When measured against self-determination theory, availability consistently substitutes for relatedness, which could fail the test, even when each conversation helps (Ryan & Deci, 2017). Availability, therefore, cannot be judged by access alone; it must also be judged by where access leads. In practice, this means building prompts toward human contact into extended or repeated sessions and reviewing aggregate patterns of use for signs of habitual reliance, as Table 1 implies.

ANONYMITY

Anonymity lowers the cost of disclosure. Where admitting difficulty or psychological stress may threaten a student's standing among peers and family, speaking to a system that does not know one's identity can be the first step a student is willing to take (Day, 2023). The same property can separate disclosure from a person who could act on what is said. A counselor who hears an indirect expression of hopelessness can ask a follow-up question, consult a colleague, or intervene. An anonymous system can do only what it has been built to do, and current systems have shown limitations in reliably distinguishing intermediate levels of risk (McBain et al., 2025). Anonymity can also make informed consent more difficult, because a student disclosing distress to an unnamed system may not know what will happen to that disclosure. The problem is not that anonymity is harmful in itself but that it may disconnect disclosure from response, which is precisely why the escalation route should be designed and tested before a system is launched. The governance response is to ensure that a route from disclosure to a person exists and works, whether or not the student has chosen to be identified.

HUMAN-LIKE WARMTH

Warmth is what makes a conversational agent feel supportive rather than mechanical. In the synthesis by Li et al. (2023), users' experience was shaped largely by the quality of the human–AI therapeutic relationship, and social presence was the explicit design principle in the course-embedded chatbots studied by Hew et al. (2023). Social response theory, however, suggests that the cues that produce this engagement can also contribute to trust that exceeds the system's competence (Nass & Moon, 2000), and that human-like cues have been shown to increase compliance with a chatbot's requests (Adam et al., 2021).

In an emotional context, misplaced trust could mean that a student may accept incorrect reassurance, postpone help-seeking because the system seemed to understand, or come to feel, as some Replika users did, that the relationship itself makes demands (Laestadius et al., 2024). As noted earlier, evidence from Replika is indirect and should not be generalized automatically to university-provided systems, but it illustrates the possibility that human-

like interaction can become emotionally consequential. Because social responses operate even when users know they are dealing with software (Nass & Moon, 2000), a single disclosure at sign-up may not be sufficient to maintain appropriate expectations over repeated interactions. Disclosure should therefore recur, and institutions should avoid design features that simulate the system's own emotional needs, since these are the features associated with concerns about dependence in the companion chatbot evidence.

DATA-DRIVEN SCALE

Scale allows an institution to extend baseline support to every student at a marginal cost that human services cannot match, although evidence from large-scale educational deployments should not be treated as direct evidence of the effectiveness of emotional support. The Georgia State deployment, for example, demonstrated this capacity for bounded enrollment support (Page & Gehlbach, 2017). This matters for equity because treatment gaps fall most heavily on groups already less likely to receive care (Lipson et al., 2022). Scale also carries two potential costs. The first concerns privacy. Emotional support systems may collect detailed accounts of symptoms, relationships, and identity, and formal work on privacy demonstrates cumulative privacy risks associated with repeated analysis and use of personal data (Dwork & Roth, 2014). For AI emotional-support systems, this supports attention not only to access controls but also to data minimization, retention limits, and restrictions on secondary use. Reviews of large language models in education already identify insufficient attention to privacy as a recurring weakness (Yan et al., 2024).

The second cost concerns bias. Expressions of distress vary across cultures and languages (Day, 2023), and a model that misreads one group's expressions may perform less effectively for students in that group. Scale, therefore, can amplify a model's limitations across a large student population, even as aggregate satisfaction remains high. The corresponding safeguards are contractual limits on retention and secondary use, a right to audit, and the reporting of evaluation results separately for different student groups.

DESIGN PROPERTIES LIMITATION

The claim is analytic rather than statistical, as it rests on how each property operates, and the evidence cited shows that each mechanism can occur, rather than establishing how often it does so. Establishing that frequency is part of the research agenda set out later in the paper. Each pairing above can be softened by design. Handoff prompts can interrupt long sessions, disclosure notices can remind students that they are speaking to a machine, and data collection can be minimized. Some improvements, such as more accurate risk detection, may also reduce harm without reducing benefit. These safeguards, however, do not necessarily affect benefits and risks in the same way. Some may reduce risk with little loss of benefit, while others may introduce trade-offs. A system that interrupts too often may feel less available, one that constantly reminds users that it is a machine may feel less warm, and one that retains little data may personalize less.

Even when design improves both sides at once, a residual risk may remain, and someone must manage that risk. Responsibility for managing that risk cannot rest with the system itself. Decisions about when a conversation should end, when a disclosure should reach a person, and what data may be kept are institutional decisions, and they should be allocated to people who can be held accountable for them (Holmes et al., 2022; Miao & Holmes, 2023). The central limitation, therefore, is not that every safeguard necessarily reduces a benefit, but that technical design alone cannot determine the appropriate boundaries of use or allocate institutional accountability. This is the governance decision to which the recommendations respond.

ADDRESSING OPPOSING VIEWS

A position is only as strong as its answer to the best objections against it. This section addresses five opposing views and explains how each challenges, qualifies, or helps define the boundaries of the position advanced her.

BETTER TECHNOLOGY WILL RESOLVE THE RISKS

The first objection is that the risks are temporary. Newer generative systems are more capable, and in the meta-analysis by Li et al. (2023), generative agents produced larger effects on distress than retrieval-based agents. On this view, governance is a stopgap until design catches up. The reply is that greater capability may improve benefits without necessarily resolving all corresponding risks. A more fluent and responsive system is also a more human-like one, which, by the logic of social response theory, may increase social responses, trust, and attachment (Nass & Moon, 2000). Crisis recognition in current general-purpose models has also shown limitations in the ambiguous middle range, where judgment matters most (McBain et al., 2025). Better technology may reduce risks, but it does not eliminate the need for someone to decide what the system is permitted to do. Some governance questions, such as who is accountable, what uses are appropriate, and when human review is required, remain institutional questions even as technical performance improves.

STUDENT SEEKING ALTERNATIVES

The second objection is that students will use commercial chatbots regardless of what universities decide, so institutional rules govern only a fraction of real use. The premise is reasonable, and it does not eliminate the case for governance. When a university licenses, endorses, or embeds a tool, it lends that tool its authority, and students may reasonably infer that the tool is appropriate for the purposes for which it is offered. That inference creates a different level of institutional responsibility (Holmes et al., 2022). Where students use tools that the institution does not control, governance takes a different form: AI literacy that helps students recognize the limits of these systems and clear signposting toward human services (Miao & Holmes, 2023). Thus, institutional governance cannot control all student AI use, but it can distinguish between tools the university provides or endorses and tools students independently choose.

PRECAUTION POLICIES

The third objection is that caution has a cost. A student on a waitlist gains nothing from a policy debate, and the realistic alternative to an AI tool is often a delay, a peer, or nothing at all. This objection highlights an important comparison, which is why this paper argues for bounded adoption rather than prohibition. The tiered model set out below proposes deployment first in areas most closely aligned with the existing evidence, namely structured self-help for students experiencing ordinary academic and transitional distress (Fitzpatrick et al., 2017; Li et al., 2023; Liu et al., 2022), while limiting autonomous AI involvement in the region where evidence is weakest. Precautions are applied to the scope of use, not to access itself. In this sense, precaution and access are not necessarily competing goals; the proposed model seeks to preserve access while imposing stronger human safeguards on higher-risk uses.

FINANCIAL CONSTRAINTS

If AI can deliver effective low-intensity support, an institution under budget pressure might reasonably consider reducing its counseling staff. The stepped-care literature, the model most often invoked to justify low-intensity interventions, addresses this objection.

The benefits of stepped care depend on assumptions about equivalent clinical outcomes, efficient use of resources, and the acceptability of minimal interventions for those who receive them; the model presupposes that more intensive steps remain available when a lower step is insufficient (Bower & Gilbody, 2005). One study in the review by Li et al. (2023) also found that participants with a more severe need for psychological support preferred human support. An institution that adopts AI-based available and support and then reduces counseling staff to the point that higher-intensity human care is no longer adequately available would move away from the supplementary logic of the model and toward replacement. Cost savings may be a legitimate institutional consideration, but they should not be treated as evidence that AI and human services are interchangeable.

GOVERNANCE IN HIGHER EDUCATION

Higher education students are adults, and they are entitled to choose where they seek support, including from a machine. The model proposed here accepts that premise. It does not restrict what students may do with tools they choose for themselves; it governs what the institution does with tools it chooses to provide. Tier 1 is opt-in, and students may move to human support at any point simply by asking. Respect for autonomy, one of the five principles identified by Floridi and Cowls (2019), can therefore support governance rather than conflict with it, because meaningful autonomous choice depends on adequate information. A student may not be able to make a fully informed choice about a system whose limits, data practices, and escalation rules have never been stated. Self-determination theory makes a related point from the student's side: autonomy-supportive environments offer meaningful choice together with a clear rationale (Ryan & Deci, 2017), rather than leaving students to discover a system's limits at the moment those limits matter most.

RECOMMENDATIONS: A TIERED MODEL OF SUPPLEMENTARY SUPPORT

The proposed tiers are conceptually derived from stepped-care principles and the evidence reviewed above; they are not presented as an empirically validated implementation framework. The governance decision identified in the central argument could be operationalized through a tiered model that matches the level of support to the level of need, adapted from the logic of stepped care (Bower & Gilbody, 2005). Table 2 sets out the proposed model. For each tier, it specifies the criteria for placement, the required actions, the responsible role, and how a case moves to the next tier. Crisis response appears as an exit from the automated pathway rather than as a further tier, because a signal of meaningful risk should trigger movement out of the automated pathway and into human review.

Table 2. *A tiered Model of Supplementary AI Emotional Support*

Tier	Triggering criteria	Required actions	Responsible	Movement between tiers
Tier 1: AI-guided self-help	Ordinary academic or transitional stress; no identified risk indicators; student opts in after limits are explained	Psychoeducation, structured coping exercises, goal setting, and reflection prompts; recurring reminders that the student is using an AI system	Director of counseling (clinical content); student affairs (deployment)	Up to Tier 2 on request, persistent or worsening distress across sessions, or a need outside scope; crisis exit on a meaningful risk signal
Tier 2: AI-assisted	Need for a service the AI	AI connects the student to the right	Named coordinator in	Up to Tier 3 when a

Tier	Triggering criteria	Required actions	Responsible	Movement between tiers
navigation and handoff	cannot provide, or distress that persists	campus service and confirms contact; it does not try to resolve the difficulty itself	the receiving service, who confirms contact within a set time	clinician judges need, or the student asks for a person; back to Tier 1 when the service needs are met, and the student chooses
Tier 3: Human professional care	Clinical assessment indicates need; student requests a person; any case escalated from Tiers 1–2	Assessment and treatment by licensed staff; AI is used only as an adjunct, with clinician approval	Licensed counselor or clinician	Step down to Tier 2 or Tier 1 at the clinical team's discretion, with the student's agreement.
Crisis exit (outside the tiers)	Any direct or indirect signal suggesting risk to self or others, including metaphorical or culturally specific expressions; threshold deliberately wide	Automated support stops; transfer or connected to a human crisis responder; out-of-hours fallback and emergency contacts displayed.	On-call crisis responder; protocol owned by the director of counseling	Returns to Tier 3 after crisis response; return to automated support only after appropriate human or clinical review

HOW THE MODEL RESOLVES THE GOVERNANCE DECISION

The model addresses the governance decision by assigning each risk in Table 1 to a named human role. Availability is bounded because Tier 1 is limited to the purposes where the evidence of benefit is strongest, and because persistent or worsening use is itself a trigger for review. Anonymity is paired with an escalation route that functions regardless of whether the student has been identified. Warmth is bounded by disclosure that recurs rather than appearing once. Scale is governed through contract, audit, and disaggregated evaluation. The model does not assume that the AI will recognize every risk. It proposes an escalation threshold deliberately set toward caution, accepting some unnecessary escalations in order to reduce the likelihood of missing more serious cases.

A brief illustration shows how a case might move through the model. A first-year student feeling overwhelmed before midterm examinations opens the university's support tool late at night. At Tier 1, the tool offers a structured exercise for managing worry and reminds the student that it is an automated system. Over the next two weeks, the student returns repeatedly and reports that sleep and attendance are worsening. That pattern meets a Tier 2 trigger, and the tool connects the student with the counseling center and confirms an appointment, which a named coordinator follows up within the agreed time. At Tier 3, a counselor assesses the student, develops a plan, and determines whether continued use of the tool is helpful. Had the student at any point written that there seemed no point in carrying on, the conversation would, under the proposed model, leave the automated pathway at once and move to the on-call crisis responder.

ROLES AND RESPONSIBILITIES

Governance depends on named roles rather than general commitments. Senior leadership, typically the dean of student affairs or an equivalent, would own the decision to deploy and would remain accountable for it. The director of counseling would oversee the clinical content of Tier 1, the escalation criteria, and the crisis protocol. Information technology and procurement staff would be responsible for data security, vendor contracts, and audit arrangements. Faculty who embed chatbots in their courses would be responsible for that decision and for knowing the escalation route. The oversight committee would oversee the evaluation and hold the authority to narrow or withdraw a tier. Students, through representation on that committee, should have a voice in how the system is described to them and in whether it is serving them. Where any of these roles is left unassigned, the corresponding risk in Table 1 may remain insufficiently managed.

PRECONDITIONS FOR DEPLOYMENT

Three conditions should be met before any tier goes live. First, an interdisciplinary oversight committee comprising representatives from counseling, student affairs, legal, information security, and students should approve the scope of use and receive evaluation findings (Holmes et al., 2022; Miao & Holmes, 2023). Second, the crisis exit should be specified in advance: the indicators that trigger it, the human role that receives it, the maximum acceptable response time, and the fallback when that role is unavailable. It should then be tested against realistic scenarios, including indirect, metaphorical, and culturally variable expressions of distress, and its failures should be documented rather than discarded. A pathway that exists on paper but has never been exercised should not be assumed to function reliably in practice. Third, because most institutions will likely license rather than build these systems, the limits on data use should be established in the vendor contract rather than in internal guidance alone. The contract should define retention periods, prohibit the use of support conversations for unrelated analytics or model training, and secure a right to audit performance. An institution that cannot audit a system it has licensed has substantially less capacity to govern its performance and risks.

IMPLEMENTATION SEQUENCE

The two campus cases reviewed earlier suggest a practical order of adoption. Both began with tasks that were high in volume and low in emotional stakes, and both kept a human professional available for anything beyond that scope (Hew et al., 2023; Page & Gehlbach, 2017). Although neither case directly tested AI emotional support, they provide useful examples of bounded institutional deployment. Institutions could therefore consider a similar sequence. A reasonable sequence begins with Tier 2 navigation, in which the system connects students to services without interpreting their feelings, and moves to Tier 1 self-help only after escalation routines have been exercised and staff have gained confidence in them. An institution that begins with an open-ended emotional-support application would be attempting to manage its highest-risk use before gaining experience with more bounded applications. Each phase should be time-limited, evaluated against the criteria set out below, and extended only by decision of the oversight committee.

CONSENT, DISCLOSURE, AND DATA GOVERNANCE

Consent should state what data are collected, how long they are kept, who can access them, and what the system cannot do. Because social responses to human-like systems persist even when users know they are dealing with software (Nass & Moon, 2000), disclosure should recur, particularly during extended sessions and at moments of emotional intensity, rather than appearing once on a sign-up screen. Data governance should follow

the principles of minimization and purpose limitation, with audit trails and restricted administrative access, in line with international guidance that calls for protecting users' data privacy in educational uses of generative AI (Miao & Holmes, 2023). Retention limits should be treated as one important privacy safeguard, alongside access controls, data minimization, and restrictions on secondary use, given the cumulative privacy risks associated with repeated analysis and use of the same data (Dwork & Roth, 2014). Consent materials should be written in plain language, offered in the languages students actually use, and revisited when the system or its data practices change, so that consent remains a continuing understanding rather than a single click.

TRAINING AND PROFESSIONAL INTEGRATION

Counseling, advising, and teaching staff need training on what these systems can and cannot do, including the capacity of generative AI to produce fluent yet inaccurate content (Miao & Holmes, 2023), so they can guide students appropriately and recognize when reliance may be a concern. Faculty who embed chatbots in coursework, as in the goal-setting and learning-companion designs studied by Hew et al. (2023), should know the escalation route as well as counseling staff do, since distress may surface first in academic settings. Staff should also be able to explain the tiered model to students in plain language, because students who understand what the tool is for are better placed to use it within its intended boundaries.

CONTINUOUS EVALUATION

Evaluation should be continuous, and it should measure what matters rather than what is easy to count. Short-term symptom scores and satisfaction ratings alone do not answer the question posed by self-determination theory: whether students become better able to manage distress independently, more confident, and more connected to others over time (Ryan & Deci, 2017). Using self-determination theory as an interpretive framework, evaluation should therefore track, at a minimum, symptom and well-being outcomes over a full academic year, patterns of use that may indicate dependency, the accuracy and speed of escalation, and outcomes disaggregated by student group, since aggregate satisfaction may obscure differences in outcomes across populations (Lipson et al., 2022).

Findings should be submitted to the oversight committee and used to adjust the scope of each tier, including withdrawing a tier if the evidence warrants it. Useful indicators include the proportion of students who meet a Tier 2 or Tier 3 trigger and actually reach a person, the time from an escalation trigger to human contact, the share of repeat users whose frequency of use increases over the semester, and differences in outcomes between student groups. Each indicator corresponds to a risk in Table 1, thereby keeping the evaluation tied to the argument that justified deployment in the first place

LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

As a position paper, this paper synthesizes existing evidence rather than generating new data. That evidence is concentrated in short trials, single sites, and a small number of national contexts. The tiered model is derived from this evidence and the logic of stepped care, but it has not been tested itself. It should therefore be understood as a conceptual governance proposal rather than an empirically validated implementation framework. Its value should ultimately be judged by whether it can be implemented and evaluated in practice, not by its internal coherence alone. The paper also does not analyze the legal and regulatory obligations governing student health information, which differ across jurisdictions and institution types; institutions applying the model will need to align their data provisions with the requirements that apply to them.

Future research should include longitudinal studies spanning at least a full academic year to track dependency formation, help-seeking, and the three capacities identified by self-determination theory, since current trials rarely report follow-up (Li et al., 2023). Effectiveness and safety should be examined across diverse student populations, including first-generation, international, racially minoritized, and neurodivergent students, with attention to how expressions of distress may vary across cultures and languages (Day, 2023; Lipson et al., 2022). Safety evaluation should move beyond single-query testing toward realistic, multi-turn conversations that include indirect expressions of risk, building on evaluations such as McBain et al. (2025). Comparative studies should also consider benchmarking AI support not only against a no-treatment control but against the realistic alternative available to the student, which may be a waitlist, a peer, or nothing at all. Implementation research should then compare institutions that adopt tiered governance with those that do not, examining both student well-being and institutional outcomes, potentially including retention and completion, which prior research has linked to student mental health (Eisenberg et al., 2009), as well as enrollment outcomes, which a bounded conversational assistant has been shown to improve (Page & Gehlbach, 2017).

CONCLUSION

This paper has argued that the benefits of AI emotional support and its most serious challenges can stem from the same four design properties: constant availability, anonymity, human-like warmth, and data-driven scale. The evidence reviewed supports the benefit side most clearly for short-term reductions in depressive symptoms and distress among young adults and students experiencing relatively manageable levels of difficulty, and it leaves the risk side, including dependency, crisis response, and equitable performance, less well established. Because benefit and risk may arise from the same properties and cannot always be fully separated through design alone, they must also be managed through governance, and governance is an institutional responsibility rather than solely a technical one.

The position advanced here is neither enthusiasm nor refusal. Universities should adopt AI emotional support only as a bounded supplement to human care, organized in tiers with explicit triggers, actions, responsible roles, and escalation routes, and with human professionals accountable for decisions involving clinical judgment, safety, escalation, or significant student well-being. Institutions that meet the preconditions set out here could offer students immediate, low-barrier support in uses most closely aligned with the existing evidence, while protecting them where evidence remains limited. Institutions that adopt these tools without such governance, or that use them primarily as substitutes for adequate human services, may shift important risks onto the students rather than managing them institutionally. For institutions considering AI in student support, the central question is therefore not simply whether the technology can provide value, but whether it can be governed deliberately enough to preserve that value while maintaining appropriate human accountability where the evidence remains uncertain.

REFERENCES

Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets, 31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: Access, effectiveness and efficiency. Narrative literature review. *The British Journal of Psychiatry, 186*(1), 11–17. https://doi.org/10.1192/bjp.186.1.11

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational*

Technology in Higher Education, 20, Article 43. https://doi.org/10.1186/s41239-023-00411-8

Day, M. J. (2023). Towards ethical artificial intelligence in universities: ChatGPT, culture, and mental health stigmas in Asian higher education post COVID-19. *Journal of Technology in Counselor Education and Supervision, 4*(1), 42–55. https://doi.org/10.61888/2692-4129.1097

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental health and academic success in college. *The B.E. Journal of Economic Analysis & Policy, 9*(1), Article 40. https://doi.org/10.2202/1935-1682.2191

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), Article e19. https://doi.org/10.2196/mental.7785

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), Article e64. https://doi.org/10.2196/mental.9782

Guo, S., Zheng, Y., & Zhai, X. (2024). Artificial intelligence in education research during 2013–2023: A review based on bibliometric analysis. *Education and Information Technologies, 29*(13), 16387–16409. https://doi.org/10.1007/s10639-024-12491-8

Hew, K. F., Huang, W., Du, J., & Jia, C. (2023). Using chatbots to support student goal setting and social presence in fully online activities: Learner engagement and perceptions. *Journal of Computing in Higher Education, 35*(1), 40–68. https://doi.org/10.1007/s12528-022-09338-x

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society, 26*(10), 5923–5941. https://doi.org/10.1177/14614448221142007

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*, Article 236. https://doi.org/10.1038/s41746-023-00979-5

Lipson, S. K., Zhou, S., Abelson, S., Heinze, J., Jirsa, M., Morigney, J., Patterson, A., Singh, M., & Eisenberg, D. (2022). Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national Healthy Minds Study, 2013–2021. *Journal of Affective Disorders, 306*, 138–147. https://doi.org/10.1016/j.jad.2022.03.038

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions, 27*, Article 100495. https://doi.org/10.1016/j.invent.2022.100495

McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Burnett, A., Kofner, A., Breslau, J., Stein, B. D., Mehrotra, A., & Yu, H. (2025). Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. *Psychiatric Services, 76*(11), 944–950. https://doi.org/10.1176/appi.ps.20250086

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies, 27*(6), 7893–7925. https://doi.org/10.1007/s10639-022-10925-9

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open, 3*(4). https://doi.org/10.1177/2332858417749220

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence, 4*, Article 654924. https://doi.org/10.3389/frai.2021.654924

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0