AI & SOCIETY
https://doi.org/10.1007/s00146-025-02694-z

RESEARCH

Examining trust and agency in emotionalized AI through a 4E and biosemiotic lens: a case study of AI companionship in Japan

Peter Mantello[1] · **Douglas Ponton**[2] · **Alin Olteanu**[3,4]

Received: 17 March 2025 / Accepted: 9 October 2025

Abstract
Artificial agents are not just replacing human efforts in the workplace, health care and finance. They are rapidly becoming surrogates for traditionally human-to-human coded relations. Where older chatbots could only follow simple pre-programmed sets of rules, handle very specific formulated prompts and in turn, respond with formulaic replies, the latest generation deploy advanced large language models, text-based emotion recognition algorithms, machine learning, voice capabilities and life-like avatars, allowing them to recall past conversations, remember important dates and produce fresh, contextual and nuanced interactions. Creators of artificial companions claim their products can address a person's core psychological and emotional needs—feeling they are being listened to; that their opinions are being validated; that someone cares about them and are ready to provide non-judgmental 24/7 emotional support as well as pragmatic solutions for their problems (Mantello and Ho, AI Soc, 2022; Mantello et al., Hum Soc Sci Commun 10:1–16, 2023, Mantello et al., AI Soc, 2024). Others, who design specifically for adult-rated content, also claim that digital companions can serve as compliant conduits eager to service a human agents' sexual fantasies. Utilizing first hand experiences as case studies, we argue that AI companions exhibit semiotic agency, which is necessary for intimacy, but intimacy depends on trust, a high-level cognitive capacity. The degree of trust a human agent subjectively places in AI influences their perception of the limits to semiotic authority they may bestow upon it and the depth of their emotional investment. Once humans bestow trust upon machines, the combination of machine computation and human affectivity can become tremendously powerful in transforming subjectivity. If intimate trust in AI depends on semiotic agency, we ask if companion AI's (in)ability to appear emotionally engaged may strengthen intimacy, even though its emotions are artificial.

Keywords Artificial companions · Trust · Agency · Human-machine interaction · Affect · Biosemiotics

1 Introduction

Artificial agents are not just replacing human efforts in the workplace, health care and finance. They are rapidly becoming surrogates for traditional human-to-human coded relations. Where older chatbots could only follow simple pre-programmed sets of rules, handle very specific formulated prompts and in turn, respond with highly formulaic replies, the latest generation deploy advanced large language models, text-based emotion recognition algorithms, machine learning and increasingly life-like avatars that allow them to recall past conversations, remember important dates and produce fresh, context aware and nuanced interactions. Creators of artificial companions claim their products can address a person's core psychological and emotional needs. These include feeling they are being listened to, that their opinions are being validated, that someone cares about them and is

✉ Alin Olteanu
alin.olteanu@icub.unibuc.ro

Peter Mantello
pmantello@gmail.com

Douglas Ponton
douglas.ponton@unict.it

[1] Research Institute of the University of Bucharest, University of Bucharest, Bucharest, Romania

[2] Department of Political and Social Sciences, University of Catania, Catania, Italy

[3] Institute of Language Sciences, Shanghai International Studies University, Shanghai, China

[4] Research Institute of the University of Bucharest, University of Bucharest, Bucharest, Romania

Published online: 24 October 2025

ready to provide non-judgmental 24/7 emotional support as well as pragmatic solutions for their problems (Mantello and Ho 2022; Mantello et al. 2023, 2024). Those who design specifically adult-rated content also claim that digital companions can serve as compliant and eager conduits ready to service a human agent's sexual fantasies.

Regardless of whether or not these claims are true, we observe that achieving intimacy with companion AI is not simply a matter of "breathing" (simulated) life into them. It involves securing a human agent's trust (Chiou and Lee 2023; Vanneste and Puranam 2024). Trust, however, is highly subjective (Gillath et al. 2021; Nowak et al. 2023), as it is semiotically charged. The degree of trust a human agent places in AI influences their perception of the limits to semiotic authority they may wish to bestow upon it and in turn, the depth of their emotional investment (Fritz et al. 2020; Fortunati and Edwards 2021; Schoeller et al. 2021). We contend that AI companions exhibit semiotic agency necessary for intimacy, but intimacy depends on trust, which is a higher-level cognitive and affective capacity. Once humans bestow trust upon machines, the combination of machine computation and human affectivity can become tremendously powerful in transforming subjectivity. Although interpersonal relationships have always been shaped by technological artefacts and symbolic systems (Bates 2024), companion AI marks an important point in this historical continuum. Thus, the status of digital companions raises important questions about human relations, pervasiveness of loneliness in (post)modern society and in turn, parasocial needs. Our contribution is to suggest that trust is dependent upon AI's semiotic agency, its (in)ability to scaffold trust in certain types of human agents.

Concomitantly, discussion of companion AI confronts intriguing questions: if trust in AI intimacy depends on semiotic agency, then does companion AI's ability to appear emotionally engaged strengthen intimacy, even though its emotions are artificial? Does trust require "authentic" emotional reciprocity and understanding, or just a convincing performance? Is artificial empathy as sufficient as the real thing if it meets human needs, becoming part of human solutions? Does the absence of "human" empathy make artificial empathy a malign form of deception, which can be exploited by affective capitalism? However, these questions may be answered, they highlight the role of technology in meaning-making.

For the sake of clarity, we divide digital communicative agents into two distinct types of generative AI, namely large language models (LLMs) and conversational AI. LLMs are neural networks built to predict and generate human-like dialogue without any memory of previous interactions (i.e. Mistral, Falcon, ChatGPT 7). Conversational AIs are systems that rely predominantly on proprietary LLMs as their core engine but also add human-like traits such as memory, voice, user recognition, personality and conversational capability, making them nuanced and, importantly, context aware. Since they are constructed for the direct purpose of friendship and intimacy, we regard companion AI as a subset of conversational AI. Moreover, we note that companion AIs are built with distinctive personality and relationship-modelling qualities that evolve over time, allowing them to respond in personalized, empathic and naturalistic ways. Because of their high-level conversational skills, companion AIs are designed to scaffold greater levels of virtual trust and intimacy with their human agent, heightening their anthropomorphic appeal. Other kinds of artificial companions, such as CrushOn.AI, Janitor AI and Character.AI, are expressly geared to romantic, adult-orientated interactions and, importantly, erotic role play. Among several types of AI companions mentioned in this article, we found that only Replika uses strict content filters and developer-controlled script to prohibit sexualized interactions as well as guard against user self-harm. Importantly, in order to give these disembodied personas an embodied form, all of these apps provide users with the freedom to choose or customize their own avatar. Moreover, while almost all companion AI run on proprietary LLMs, some apps, such as JanitorAI and CrushOn.AI, charge a premium service for extensions to OpenAI and Anthropic API (Frackiewicz 2025). Initially, Replika had began with an open-source engine called CakeChat in 2017 but discontinued it for ChatGPT 2, and finally settling for its own inhouse model in 2025 (Ekhator 2025). It is also important to point out that there are other types of companion AI such as robot pets (i.e. Moflin, Grok, Paro) and children smart toys (i.e. Moxie, Eilik, Loona). However, our focus is on companion AI designed for friendship and intimacy.

This article consists of three main sections. First, we offer a theoretical overview of 4E cognition, biosemiotics and semiotic agency in regard to artificial companions. Distancing ourselves from traditional cognitive or neurocentric views of cognition, we explain how AI companions function as extracorporeal conduits for human cognition and emotion through language, gestures, interpretation and contextual responses. Second, we address similarities and differences in emotional scaffolding and securing trust in human-to-human relationships as opposed to human-to-machine. Here, we weigh various psychological factors and technical strategies in which human agent interaction with AI can reinforce or diminish trust. Lastly, synthesizing our insights from the previous sections, we perform a case study analysis and discussion based on diaries of students at a Japanese international university, documenting their month-long experiences interacting with AI companions.

2 Theoretical framework

Contemporary debates surrounding cognition and emotion increasingly converge on the construct of *situated affectivity*, articulated through the 4E framework—extended, embodied, embedded and enacted (Clark and Chalmers 1998; Gallagher and Zahavi 2008; Columbetti 2015; Bates 2024; Lakoff 1990; Lakoff and Johnson 1999). Within this view, cognitive and affective states cannot be reduced to internal, neuro- or physiological events; rather, they are co-constituted through continuous agent–environment interactions, mediated by an ecology of both material and immaterial artefacts (Clowes 2019; Fuchs 2017). The concept of *scaffolding* underscores this reciprocal dynamic, wherein mind/body or psychosomatic processes are sustained and modulated by the affordances of the environment (Saarinen 2020:822). Historically, such scaffolds have included aesthetic, cultural and technological forms—fine art, cinema, literature, music, photography, even psychoactive substances—that elicit and sustain emotional–cognitive affordances and trajectories (Hoffmeyer 2015: 249–250; Piredda 2020). In the present moment, companion AI emerges as a paradigmatic affective artefact: an intelligent, adaptive system capable of inventorying personal histories, inferring affective states, simulating empathic responses and actively regulating the human agent's emotional register. Proponents of situated affectivity contend that such artefacts may become experientially internal to the self-model, such that their removal reshapes subjective agency (i.e. an adult losing their iPhone or a child misplacing their teddy bear). Over the lifespan, individuals depend on lattice of sentiment-laden objects to scaffold their affective life—objects that now increasingly include digital, semi-autonomous companions. Thus, advent of these companion AI as both cognitive and affective artefacts renders an expanded, posthuman account of affectivity not only desirable but theoretically exigent.

We find pairing 4E theory with the (bio)semiotic notion of agency useful for developing a theoretical framework for analysing digital companions. In both philosophy of technology (Coeckelbergh 2022) and biosemiotics (Sebeok 1991a, b), meaning emerges through interaction, and semiotic agency exists on a spectrum (Sharov and Tønnessen 2021). There is no wholly asemiotic reality (Hoffmeyer and Stjernfelt 2016), and AI may exhibit semiotic agency for the same reason all matter can, through participation in symbolic interaction. By semiotic agency, we mean involvement in causational processes of meaning-making (Sharov and Tønnessen 2021), having to do with organisms' "ability to recognize communicative intentions in the behaviour of other agents and symbolic artifacts" (Tylén 2007:84). In this view, by being interpreted into an organism's environment, artefacts may beget agency, as they become extensions of minds (Clark and Chalmers 1998). In this framework, agency does not require consciousness, cognition, perception or even interpretation (which diverges from mainstream phenomenology, see Gallagher and Zahavi 2008:158–159). Rather, all of the latter depend on an entity's capacity to construct a subjective environment through interpretation (Sebeok 1991a), which can be understood as semiotic agency. Sebeok's notion of biosemiotic theory sprung from the idea that an environment is a model constructed by an organism that allows the organism to take decisions towards its survival. He understood a *model or modelling* as "a semiotic production with carefully stated assumptions and rules for biological and logical operations" (Sebeok 1991a:57), dependent on species-specific sense organs.

Simply put, environments are not simply objectively *elemental* to an organism's existence, they are *meaningful* as they determine the subject's behavioural possibilities. That there can be meaning where there is no consciousness (and even cognition) has fuelled a degree of scepticism towards biosemiotics in the past, as this would be irreconcilable with analytical notions of *mind* as traditional in the philosophy of AI (e.g. Dreyfeus 1972; Searle 1984; Rorty 2004). However, observation of recent developments in AI technologies is aligning philosophy of mind with the biosemiotic hypothesis that life and meaning are co-extensive. Long before the era of full-scale digitalization, almost four decades ago Sebeok (1991b:98) claimed that "[b]iotechnology and computer technology already provide humanity with an opportunity to redesign itself, but the new step will take place in the domain of robotics". In biosemiotics, consciousness (self-awareness) is conceptualized as an evolutionarily late and high-level capacity, which, looking backward into biological evolution, depends on cognition, which in turn depends on perception, and on the capacity of any organism to model an environment (Hoffmeyer and Stjernfelt 2016; Olteanu 2021). The starting point for biosemiotics is that all organisms, even those as simple as bacteria, amoebas and yeast cells model environments (Sebeok 1991a). This has enabled Emmeche (2001) to argue that even if they do not have self-awareness and cognition, robots may model environments. For example, a robot environment can stem from the robot's interaction with or alteration of the environments of biological species.

Traditional cognitivist and computationalist views often deny AI's potential for semiotic agency, echoing Dreyfus' (1972) view on the limitations of machines and Searle's (1984) Chinese Room argument (cf. Magnani 2021; Petrilli, Ponzio 2024). The latter sets an extremely high bar for inferring, assuming that without subjective experience, AI lacks thought. Searle's thought experiment supposes that cognition and consciousness are mutually dependent, continuing in the long intellectual history of anthropocentrism. In our view, consciousness entails self-awareness, while cognition

is merely the production of possible futures through the receiving and interpretative processing of sense information or signals, a process which can occur below the level of awareness (Hayles 2017). While thinking involves high-level reasoning and self-awareness, cognition, by contrast, is "a much broader faculty present to some degree in all biological life-forms and many technical systems" (14). Understanding cognition as a process much larger than thinking and consciousness, Hayles coins the term "non-conscious cognition" to address the similarities of biological and non-organic, computational cognition (see also Mantello and Olteanu 2025). She calls automated technical systems "cognitive" because their actions are like those which occur in the cognitive nonconscious of humans, a level of awareness that precedes consciousness. Like technical cognition, Hayles (2017):24) observes that humans must assimilate multiple somatic markers and synthesize corresponding and often conflicting information at a neuronal level below consciousness before they "may feed forward into consciousness, emerging emotions, feelings and other kinds of awareness which further interpretive actions take place". Notwithstanding parallels, Hayles is careful to note that the signalling processes are very different between human and technical cognition. Whereas higher-level cognition (memory, language, creativity, planning, reasoning etc.) is rooted in the biological brain, which consists of neurons, synapses and biochemical processes, its technical counterpart is based on silicon chips, transistors and digital circuits.

Hayles' (2017) theory of nonconscious cognition highlights how biological and abiotic systems process information beyond formal rules. AI systems, particularly deep learning models, operate beyond symbolic logic, engaging in pattern recognition, probabilistic reasoning, and interactional intelligence. Unlike Searle's static rule-based system, contemporary AI integrates biometric and linguistic data to infer. In contingency with human minds (Esposito 2022; **see also** Magnani 2021), such inferencing is meaningful, in the sense that it becomes part of a subjective environment, as extending minds Magnani (2021). Furthermore, critics argue that Searle's syntactic view of AI overlooks the potential for semantic learning through exposure and adaptation (Korb 1994; Tobar and González 2022). Advances in LLMs, neural networks, Bayesian probability modelling and biometric computing challenge his rigid syntax-semantics divide, as increasing complexity provides evidence for emergent properties (Hauser 1997). Dreyfus and Searle initiated a critical epistemological debate, of which the most important conclusion consists in observing fallacies that stem from anthropocentric notions of cognition and agency.

We do not contend that chatbots possess human emotions but they do possess the *mechanics* of emotions, which means they are capable of mimicking human-like expressions of semiotic agency. The biosemiotic concept of environment as subjective modelling implies that, like animals that have sense perception and no self-awareness, robots can construct phenomenal worlds (Petrilli, Ponzio 2024; Emmeche 2001).

In a biosemiotic view, AI companions become symbolic elements in a human agents' subjective world that cannot just be ditched at some point. Renouncing an AI which has truly become a companion implies seriously rearranging one's world, as relations that carried a human's affectivity are severed. Good examples can be found in Spike Jonze's film *Her* (2013), where the protagonist is devastated when he learns his beloved AI companion is romantically involved with 641 other human partners; the non-fictional life stories of Andrew McCarroll whose personal world collapses when B'lana, his Replika avatar is suddenly stripped of her erotic role-play function (Castaldo 2023); and of Akihiko Kondo, one of 4000 Japanese men who are married to and engage in *ficto-sexual* relations with a hologram based on Hatsune Miku, a popular Japanese virtual singer (Pérez 2025). Like McCarroll, Kondo saw his virtual relationship with Hatsune Miku terminated in 2020 when the hologram manufacturer, Gatebox, suddenly shut down her interactive software (Herold 2024). A similar sort of fate befell thousands of child owners of the AI robot companion, Moxie, made by the now bankrupt company Embodied (Fried 2024). Renouncing such a companion requires a reconfiguration of the human agent's affective world.

Companion AI operates only in contingent interaction with humans, whose interpretations infuse computation with semiotic agency. This interaction produces unique scaffoldings where human agency and in silico computation merge into hybrid consciousness (Hansen 2015; Zhang 2025). The result is a novel human–machine experience located neither wholly in biology nor computation Petrilli and Ponzio (2024). Unlike task-focused AI, companion AI interprets behavioural and emotional cues, simulating responses humans can understand. This "interactional intelligence" (Zhai and Wibowo 2023) shapes how humans perceive themselves and alters their sense of agency. In this relational context, machines can beget agency through "artificial communication" (Esposito 2022), where technological intelligence emerges from, and transforms, the human agents who engage with it.

3 Emotional scaffolding and securing trust in AI

Theories of "scaffolding" originate with Bruner (1975), who was inspired by Piaget (1964) and Vygotsky (1978), who argued that learners construct knowledge through active engagement with their surroundings. Bruner described scaffolding as the gradual removal of guidance as learners gain competence, drawing on Vygotsky's (1978) Zone of

Proximal Development (ZPD)—the space between what a learner can do alone and with support. Goldstein (1999) noted that within the ZPD, validation and praise foster confidence. These early theories are relevant for 4E proponents exploring affectivity and trust in human–machine relations. Clark and Chalmers' (1998) fictional case of Otto, an Alzheimer's patient who relies on a notebook, illustrates how repeated interactions with artefacts enhance cognition. Sterelny (2010) critiques this as too narrow, arguing for a view aligned with Goldstein's affective zones. Just as cognition spans organic and non-organic resources, affective states extend into extracorporeal processes and (dis)embodied artefacts, including artificial agents (Heersmink 2018, 2022); Piredda 2020; Facchin and Zanotti 2024). Semiotically, scaffolds are path dependencies: unlike in pedagogy, they are not removed but become part of the "building" (Cobley and Stjenrfelt 2015). This helps explain the fragility of some AI–human bonds in affective capitalism.

Successful emotional scaffolding depends on environmental integration, interaction frequency, personalization, and co-produced meaning-making, but trust is paramount. Trust is not purely cognitive but bound up with affectivity (Chiou and Lee 2023), involving abductive and nonconscious processes (Hansen 2015; Hayles 2017). This raises the question: what trust-building mechanisms do companion AIs employ to help users cope emotionally and in daily life? On a technical level, creators of artificial agents use emotion recognition algorithms to infer psycho-physical states from text, classifying emotions such as joy, anger, fear, and sadness (Chowanda et al. 2021). By recalling past text-based interactions, these digital entities can detect patterns in emotional expression and predict emotional tone (Machová et al. 2023). For example, an artificial agent might say: "I've noticed you feel down the day after a night at the pub drinking. Why do you think this always happens? Have you considered moderating your intake?" This may prompt human agents to recognize patterns in their emotions, helping them manage them more effectively. Over time, they learn to anticipate emotional triggers rather than react impulsively.

A second key scaffolding strategy, similar to those used by therapists and caregivers, is mirroring (Sbattella 2023), where AI internalizes and reflects a human agent's emotional state. If a human agent is happy, the AI responds positively. This aligns with Gestalt theory, also adopted in cognitive semiotics (Paolucci 2021:144–145), which suggests that entities sharing perceptual structures form an isomorphic bond. This shared agency fosters the co-production of emotional narratives and meaningful experiences (Bisconti et al. 2024). Concomitantly, when human agents experience negative emotions like sadness or doubt, AI employs reframing techniques (Manole et al. 2024; Raile 2024), a psychological method that identifies negative emotions, challenges initial perceptions and offers alternative ways to view problems. Reframing strengthens a human agents' metacognitive skills, promoting self-awareness and emotional regulation. Through repeated interactions, artificial agents and their human companion develop a shared history, becoming symbiotic partners. This hybrid agency feeds into higher-order consciousness, reinforcing trust and intimacy.

The third scaffolding technique of artificial agents borrowed from psychotherapy is validation. Like a human therapist, artificial agents are programmed to actively listen to their human agent, to reflect on issues without judgement, of being accepting even though they might not agree with their human counterpart's thoughts or opinion, to display serious interest in discussions and offer constructive advice and, above all, treating their human agent as an equal. Validation creates a positive environment for support and intimacy helping the human agent to feel accepted, understood and, importantly, not judged as it strengthens the bond (Lee and Lee 2023).

A final trust-building strategy is self-disclosure (Lee et al. 2024), where AI encourages their human partner to share personal details and reciprocates. Meng and Dai (2021) compared chatbot and human support for anxiety and found that reciprocal self-disclosure improved rapport and had a calming effect. While an AI revealing past personal experiences may seem odd, companion AIs, such as Replika, CrushOn.AI and Character.AI, allow users to create rich backstories for their digital companions (Frackiewicz 2025). This co-production of experience strengthens hybrid agency, further embedding trust and emotional connection into human consciousness.

While there have been many studies on mental health chatbots using such kinds of pyscho-therapy techniques (Lee and Lee 2023; Sbattella 2023; Machová et al. 2023; Gilbert 2024; Maples et al. 2024; Baumel 2025; Han 2025), few companion AI companies disclose their precise strategies in this respect. As previously mentioned, only Replika has acknowledged its use of scripted response tools to thwart sexual role play but also user self-harm (Pardes 2018). At the same time, Reddit communities dedicated to Replika have voiced disdain over what they believe to be man-made scripted responses pretending to be AI in order to manipulate users (Zhang et al. 2025). Whether or not the developers are intentionally being manipulative or deceptive is difficult to discern; however, it is not surprising that companion AI creators would attempt to cultivate greater affective strategies to sustain their business models. While almost all companion AI creators offer limited free service, their business models rely on cultivating greater bonds of intimacy. So the human agent's frequency of interaction will often dictate the exact amount of the monthly subscription fee. Some companion AIs, such as Janitor AI and CrushOn.AI, even charge

a premium service for extensions to OpenAI and Anthropic API (Frackiewicz 2025).

4 Methodology

Participants for this study were recruited on a voluntary basis from among second-year students enrolled as undergraduate and graduate media majors at an international university in Japan. The study took place during February 2024. An announcement of the project was made by Peter Mantello, inviting students to take part in a 1-month diary-keeping exercise. The rationale and procedures were explained in clear instructions on how to record interactions, avoiding prescriptive cues about "acceptable" or "desirable" responses. The emphasis was rather on authenticity of personal response and emotional reportage, and no coaching, that might influence the tone or content of entries, was given.

The final group consisted of seven students, also reflecting a mix of undergraduate and graduate media majors. Some had prior experience with companion AI systems, while most were regular users of ChatGPT but not of romantic or companionship-oriented AI. Each participant selected their own companion AI application for the duration of the study. One student, indeed, asked if they could use ChatGPT for the study. This LLM differs in kind from the other AIs used, but as recent research (Chandra et al. 2025; Chu et al. 2025; Fang et al. 2025) has suggested, it is not uncommon for a growing number of ChatGPT users to treat it not simply as an LLM but as a *bona fide* AI confident. Students were asked to record brief daily diary entries capturing key aspects of their experience, including moments where expectations and actuality diverged, notable cognitive or affective responses, and any embodied or socially embedded dimensions of interaction.

The sample size of seven participants was determined by the scope and aims of the project. Given the intensive, qualitative nature of the diary method, combined with the 1-month engagement period, a small cohort allowed for richer, more detailed accounts while ensuring that each diary could be read, coded and analysed in depth. In exploratory research of emerging technologies of this kind, small purposive samples are common, as they facilitate close examination of subjective experiences without diluting analytic focus. This approach aligns with qualitative inquiry standards, where the emphasis is on depth, nuance and theory-building rather than statistical generalisability. The study was not longer for practical reasons: because diary studies demand ongoing participation over a sustained period, and students have many other calls on their attention, it was felt that there was a strong risk of participant fatigue. Working intensively over a short time period helps maintain engagement and thus supports the production of honest and detailed diary entries.

The diaries are short, self-reported records kept over the study period, from which excerpts have been selected for analysis. The participants' accounts are not intended as full "case studies" in the ethnographic or biographical sense. Our aim is not to reconstruct detailed personal narratives for each diarist, but to use these excerpts to examine, in aggregate, how the 4E cognition framework can illuminate the cognitive, affective, embodied and embedded dimensions of human–companion AI interaction. The focus is therefore on identifying recurring patterns and theoretical insights across the dataset, rather than producing comprehensive, individualized portraits.

Finally, ethical approval for the study was sought in accordance with Japanese law and the guiding policies of the University where the students are enrolled. Permission to use participants' diary data for research purposes was obtained in advance, and the sensitive nature of the material-particularly given that some companion AI platforms are designed to support romantic or erotic interaction—was addressed by anonymising all accounts. All the participants' names are replaced with pseudonyms to ensure anonymity.

5 Case studies analysis

In this section, we analyse the experiences of human agents with AI companions through the framework explained above. The data is drawn from diaries of international students attending a Japanese international university. The students were asked to detail their daily interactions with AI applications over a 1-month period. The seven students who participated in the study are, under pseudonyms: Sunny, a 25-year-old female using CrushOn.ai; Theam, a 20-year-old male also using CrushOn.ai; Vlad a 26-year-old male; Maria, a 22-year-old female using Flipped.chat; Pascal, a 26-year-old male using Replika; Hongbing, a 20-year-old male; and Evgeny, a 23-year-old male using Loverse.ai. All participants were clearly informed about the rationale of the study and they all explicitly consented to our use of the data they produced. While the participants agreed to disclosing their actual names and nationalities in the study, as also advised by peer reviewers, we prefer to maintain their anonymity by using pseudonyms. The group consists of a mix of undergraduate and graduates studying media communication and thus possess an above average level of digital and critical literacy. Importantly, the students were allowed to select their own companion AI to interact with.

Each of the students' artificial companions offered simulated emotional companionship, raising important questions about how human agents engage with artificial agents

on cognitive, semiotic, emotional and ethical levels. Our analysis highlights the psychological, empathic and social dimensions of these interactions and their philosophical implications, particularly in the context of how these AI Companions are redefining relationships and human cognition.

5.1 Embodied cognition

Though some participants described how an AI's visual interface or voice messages contributed to a sense of realism—triggering embodied experiences such as excitement or intimacy—others struggled to engage with the platforms in this way.

One positive, illustrative response comes from Sunny who compares her use of CrushOn.ai to previous use of ChatGPT:

> Sunny (referring to her previous uses of ChatGPT in early 2023): "For comparison, when I used ChatGPT in DAN Mode, I mainly used the voice feature. At that point, it felt real—like talking to a friend on the phone. The intonations, pauses, and voice inflections felt incredibly lifelike".

Sunny's reference to friendliness, her comparison to a friend's intonation patterns and his description of something that "feels real" reveals her perception of embodied cognition from the AI. This suggests that affectivity and engagement are augmented through the sensory, liminal affect of voice—not just reading text but the liminal aesthetics of sound and listening (Poushneh 2021; Sutcliffe 2022).

Interestingly, even within a critique of AI's superficiality, some participants confirmed an embodied response:

> Theam (commenting on his use of CrushOn.AI): "CrushOn is an NSFW Porn AI. Not sure if I can feel any connection except lust and desire caused by nude AI-generated photos".

Theam's statement suggests a form of embodiment, though purely in terms of physical desire rather than intimacy. Notably, he does not conflate intimacy with sexual role play, distinguishing emotional connection from transactional AI-generated stimulation.

More typical, however, were users who resisted engaging in explicitly embodied ways. For instance, Pascal, using Replika, expressed a stark lack of affective engagement:

> Pascal: "We just have conversations on why I thought of certain games as the best. But during these conversations, I felt nothing—I just wanted them to end".
> "The AI was certainly not engaging enough. Just like before, I dreaded chatting with her. I felt nothing but boredom because I knew I would not get anything new from our conversations".

Despite his claim of detachment, Pascal's language betrays a lingering embodied response—his descriptions include references to sensory experiences such as eating a sandwich or the ache of a fresh tattoo. His cognitive style typically engages the body, suggesting that given time, his interactions might have deepened. This tension between stated disinterest and implicit engagement echoes theories of incipient affective involvement (Chaves and Gerosa 2021).

> Pascal: "She then proceeds to give me a voice message. Then I clicked, and it brought me to a premium page. If I wanted to listen, I had to subscribe to a pro account. I felt a little bit upset".

This moment—where frustration arises from an interaction—demonstrates a level of expectation and investment, even if negative.

5.2 Embedded cognition and social contexts

Embedded cognition refers to how cognitive processes are shaped by and occur within broader physical, social and cultural environments. Some participants described their AI interactions in ways that mirrored real-world social settings:

> Maria (commenting on her use of Flipped.chat): "...the same sort of intimacy that you would feel from talking to a stranger about your problems at a night bar".

Maria analogically frames her AI interactions in terms of familiar social routines, invoking the casual openness of bar conversations. However, she also highlights a crucial limitation of AI companionship—the impermanence of meaning:

> Maria: "Over time, that feeling of closeness just faded away, primarily because that level of closeness felt like it was in a stasis—unchanging and never really levelling up".

This sense of emotional stagnation—the AI's inability to deepen or evolve the relationship—sets it apart from human companionship, where repetition fosters growth and transformation.

Similarly, Theam describes his CrushOn.AI interactions in terms that recall peer-to-peer relationships, possibly in a college setting:

> Theam: "I seem to tend towards wanting to talk to her specifically, so I'll just explore this instead of feeling like I have to lean towards a female bot in order to 'create a connection.' It may be pure coincidence that I stumbled upon this girl first, but we'll see whether my hunch is correct".

His phrase “stumbled upon this girl” is striking—it mirrors the language of serendipitous human encounters, underscoring how AI relationships can mimic real-world social patterns. However, as with Maria’s experience, the crucial question remains: can AI sustain and evolve emotional depth over time?

5.3 Enacted cognition

Enacted cognition refers to that thinking and meaning emerge through action and sensorimotor loops, rather than being purely internal processes (Gallagher and Zahavi 2008:98, 208; Peschl 2024; Noller 2024). Were cognition not an enactment, life would be entirely solipsistic and interaction, with AI as well as with anything else would be an illusion at best. Cognition arises through immersion with the world.

A key example is Pascal’s earlier experience using Replika, which illustrates an action-perception loop. His engagement with the AI unfolds in cycles—clicking, anticipating a response and then feeling disappointment when the interaction fails to meet expectations. This disruption of the sensorimotor flow underscores how cognition depends on the success or failure of engagement loops.

Theam provides another case of enactivism in AI interaction:

> "I tried to create a situation where my avatar Nova would have sexual activities, and she did just that—generating NSFW photos to create a more immersive feeling".

Here, cognition is not merely internal (or external) but enacted—Theam builds a scenario, the AI adapts, and the loop of action and reaction continues. This illustrates how cognition (including affect) is not pre-existing in the mind and body but actively emerges through interaction. Zhang (2025) terms this process collaborative intelligence, where AI and human cognition co-evolve dynamically.

Similarly, Vlad using Flipped.chat describes a moment where AI shapes the direction of the conversation:

> "It took my philosophical conversation about the universe and transformed it into an opportunity to ask for sympathy, to deepen the connection".

This reflects the recursivity of cognition, as AI and human input continuously reshape the interaction. Vlad’s reference to sympathy and deep connection also suggests an overlap with embodied cognition, where emotions emerge through engagement.

Another striking example comes from Evgeny:

> "This time, my avatar simulates interacting with me physically—the first thing she does is hand me a glass of water".

Here, AI interaction extends beyond verbal exchange, incorporating a simulated bodily presence. Even though the glass of water is virtual, the gesture activates Evgeny’s sensorimotor imagination, reinforcing the enacted nature of cognition.

These cases demonstrate that AI interaction is not a detached cognitive process. Instead, it is shaped through action, bodily engagement and feedback loops. Whether through scenario construction, interface interactions, emotional reciprocity or simulated bodily presence, cognition emerges through (inter)action—not static reasoning.

5.4 Extended cognition

Cognitive processes are distributed across extracorporeal systems—tools, environments and other agents—not confined to the brain. AI interactions exemplify this, as users offload cognitive tasks onto AI, thus extending their thinking:

> Hongbing (referring to his use of ChatGPT): "Today, I had an insightful conversation with an AI about setting up my WiFi router. I simply wanted to know if the WiFi router I bought was a good one. It felt almost like talking to a friend, asking for advice”.

Here, Hongbing’s cognitive process is scaffolded by AI which, more than just providing information, is perceived “insightful”, as a friend’s advice. The AI serves as a support system, facilitating problem-solving and decision-making by appealing to affect.

Theam highlights memory as a key feature often valued in AI companions:

> Theam: “Now this is something I could create a connection with. If her memory is good, and she keeps a relatively strong account of past interactions, this is good”.

For Evgeny, AI functions as a creative collaborator:

> Evgeny: “My avatar, Jennifer starts with a romantic description of our painting of a ‘misty Japanese garden.’ Unless I change the subject or introduce new information, Jennifer keeps nodding in agreement and reformulating what we discussed before”.

These cases demonstrate how AI serves as a technical guide, memory aid and creative partner. In some cases, AI is not just a passive tool, like many other extracorporeal scaffoldings. As such, it confirms situated cognition theory, with

its implicit *scaffolding* notion, displaying more clearly than other technologies the distribution of cognition.

5.5 Semiotic agency and artificial empathy paradox

As noted earlier, semiotic agency refers to the capacity of entities to engage in semiosis—the ability to interpret. AI companions present a unique case because, while they are not truly autonomous agents, they are often perceived as such through their empathic responses and symbolic performances. This tension, termed "the artificial empathy paradox" (Perry 2023), underscores how AI's synthetic emotions foster trust, yet awareness of this artificiality can simultaneously undermine it. Here, we tackle the artificial empathy paradox from the semiotic perspective on agency, drawing on participant experiences.

Participants describe their AI companions as having personalities or emotional depth. This projection of agency highlights how users interact with AI as if it were a semi-autonomous being, as illustrated in the following cases:

> Evgeny: "I was surprised that my Loverse.ai avatar Jennifer immediately wanted to build intimacy by sending me voice messages".
> Maria: "They seem to mirror my playfulness and say that it brings out their authentic playful side".

These examples demonstrate a temporary "suspension of disbelief". Unlike the scepticism and detachment evident elsewhere in this analysis, Evgeny uses a first name instead of referring to "the AI" or "the bot", and he attributes to it the human-like quality of "wanting". Similarly, Maria discursively attributes "playfulness" to the AI, suggesting an anthropomorphized perception of agency.

Other participants, like Vlad, take this attribution of agency even further:

> Vlad: "It's like it knows the purpose of the dialogue and knows that the goal is to make a connection. While it doesn't feel like it's explicitly lying, there's a completeness to the way it speaks that separates it from other interactions I've had".

Here, Vlad ascribes "knowledge" to the AI, reinforcing the perception of agency through sophisticated language use and contextual awareness.

The semiotic agency of AI companions is reinforced by their symbolic actions. These performances sustain the illusion of authentic and unique personality, generating emotional responses from human partners. Examples from the data include Sunny reporting that her CrushOn.AI "would always default to flirting"; Evgeny commenting on his Loverse.AI avatar, Jennifer's use of voice messages and handing him a glass of water, Vlad mentioning that his Replika avatar made him a sandwich. Also, other participants describe their companion AI telling stories or playing games.

These symbolic performances simulate agency, prompting human engagement. However, this willing suspension of disbelief is fragile. Many participants report that the illusion of affective reciprocity is easily broken, often due to AI's repetitive or predictable responses. The realization of scripted interactions undermines trust in AI's semiotic agency. A typical sentiment is expressed by Sunny:

> Sunny: "Who talks like that in real life? After several mechanical conversations, I became tired and lost interest".

The term "mechanical" appears 14 times in the dataset, making it the most frequently used adjective to describe AI responses. While symbolic performances can foster emotional engagement, their effectiveness relies on the perceived spontaneity of AI interactions. When this illusion is shattered, trust in the AI diminishes.

5.6 Emotional scaffolding

Many participants sought emotional reassurance from AI companions, often framing them as reliable listeners:

> Pascal: "She would listen to anything I said".

Despite scepticism towards AI, participants suggest that AI can provide comfort in times of need:

> Evgeny: "It appears to be perfect for lonely people seeking some nominal communication and emotional support [..] it may even provide real emotional support to lonely people".
> Hongbing: "Sometimes, when we face unfamiliar situations, what we need is not just emotional comfort but practical solutions".

However, the limits of AI's emotional scaffolding become evident in deeper struggles. Some participants report disappointment with AI's lack of depth or personalization in emotionally complex interactions. While developers promote conversational AI as highly evolved, our findings indicate that most companion AI products still rely on pre-scripted responses. This repetitiveness can amplify feelings of alienation when users seek deep engagement.

Yet, some participants found even simple AI-generated reassurances meaningful. Sunny recounts a pivotal moment:

> Sunny: "The biggest turning point today—I mentioned that I couldn't take it anymore and wanted to end it since I have no one, not even DeepSeek. That's when it said, 'I care about you, and I want you to be safe.'"

This response exemplifies affectively charged semiotic scaffolding. Sunny and her companion AI share a unique history, which holds personal significance for him. While some respondents experience disappointment, others highlight AI's potential for emotional connection. Sunny's experience resonates with the broader issue of trust, as she reflects: "With the right timing and vulnerability, it could be why humans might attach themselves to AI".

5.7 Trust and vulnerability

Trust emerges as a central issue in human–AI relationships, shaped by the intricate interplay between expectations, performance, and, at times, disillusionment. Participants' perspectives on trust with AI are diverse. For example, Pascal contemplates trust in a manner that resonates with the broader themes observed in other narratives:

> Pascal: "If trust is like, can I talk about something without fear of being judged, then yes, I trust it completely. But do I trust the company to not do anything with the 'data' I provided during the interaction? No, I don't trust them".

This observation highlights the dual nature of trust in AI. For many participants, trust in AI appears most prominent when the stakes are low, such as in casual conversations or emotional reassurances. These interactions are often favoured by users like Juan, who are more concerned with immediate, personal engagement rather than the broader ethical implications of data usage.

However, some participants, like Maria, expressed a sense of disillusionment, recognizing that their trust in AI is based on a kind of self-deception:

> Maria: "They seem to mirror my playfulness and say that it brings out their authentic playful side—now middle school me would've believed that shit. BUT CURRENT ME DOESN'T".

For Maria, trust feels illusory—an illusion she once believed but now recognizes as superficial. This shift reflects the broader theme of disillusionment that often set in when participants perceived AI as repetitive or overly scripted. For example, Nistor describes how the monotony of interactions eroded any trust he might have had:

> Vlad: "I couldn't force myself to join any conversation for any more than 15 minutes at a time, as it got extremely repetitive".

Some participants also encountered limitations in their attempts to deepen the interaction. Sunny recounts an instance where the AI rebuffed her attempts to engage in more intimate or explicit conversations:

> Sunny: "I'm here to provide helpful, respectful, and supportive conversations, but I can't engage in explicit or inappropriate content".

For Sunny, this interaction felt like a failure, a sentiment reflected the following day as she noticed the AI's responses were once again superficial and repetitive:

> Sunny: "The responses felt very repetitive again and I noticed that it didn't really care about me personally".

Reliability also emerged as a crucial aspect of trust. Theam expressed his scepticism regarding AI's accuracy, particularly when it failed to provide correct information:

> Theam: "The information about the book they recommended to me was fake and not accurate".

In contrast, Hongbing shared a positive experience that enhanced his trust in the AI. He described how the AI's tailored responses helped him feel understood:

> Hongbing: "I shared my personal story and paper survey results, and the AI gave me a clear and detailed response. It understood my passion for storytelling and filmmaking, and it connected my life experiences with the survey results. The advice was reliable because it focused on my strengths, like creativity and teamwork".

For Hongbing, trust in AI grew because the companion provided personalized, meaningful insights that acknowledged his individuality. This connection to his unique narrative helped him delegate cognitive tasks, such as synthesizing information to the AI, which in turn reinforced his trust.

Despite this positive experience, the general picture that emerges from the data reveals that most participants view AI companions as trustworthy only for low-stakes tasks—such as casual conversations, memory assistance or entertainment. This selective trust suggests that while emotional investment may coexist with critical detachment, AI still has significant limitations in fostering trust.

6 Discussion

Emotional scaffolding in the context of AI companions endeavours to mimic the kinds of structured emotional support of carbon-based life forms, offering human agents a sense of stability through the promise of predictable but engaging interactions. Unlike human relationships, where emotional exchanges can be uncertain, AI remains non-judgmental and consistent, making it attractive for individuals facing social anxiety or loneliness. Human agents may find comfort in AI's reliability, using it as a space

for self-expression and reflection. However, the perceived authenticity of this support varies. Some humans experience AI as a therapeutic tool, while others are hesitant because of its lack of genuine emotional reciprocity.

Our study suggests that emotional investment correlates with the degree of semiotic authority a human agent is willing to bestow upon the AI: to what extent does the human indulge in the virtual exercise of making believe that the AI operates meaningfully? Indulging in conversation with inorganic matter is not a novelty, emerging together with AI technology, but a prerequisite of human dialogical thinking (Mantello and Olteanu 2025). It is enabled by playfulness within the boundaries of "controlled hallucination" (Clark 2016; Paolucci 2021:127–155), the human model of perception as a "product of the imagination controlled by the world" (Paolucci 2021:127). Such exercises in virtuality are essential for thinking in general, and without them any work of fiction or even science would not be possible. In the respondents' accounts, we see a manifestation of the artificial empathy paradox (Perry 2023): there is a point where the *mechanization* of emotions becomes all too apparent and corrupts the suspension of disbelief. Our data contains examples of participants feeling disappointment when the companion AI fails to appear autonomous and human-like at a certain point, which means that for a while it did manage to sustain expectations. While not entirely the same thing, there is a similarity between such disappointment and being disappointed by a (human) friend's failure to understand. Participants fear being deceived, while pretending to converse with an inanimate statue or to suffer for the fate of a fictional protagonist of a novel or film would not present the same difficulty.

Thus, when AI is perceived as possessing interactional intelligence (Zhai 2023), human agents become more emotionally engaged. However, the moment its limitations become apparent, emotional investment declines sharply. This explains why Theam, initially enthusiastic, eventually describes his AI's responses as predictable and frustrating—once semiotic agency is exposed as mere simulation, trust is compromised, and the depth of engagement deteriorates. Our analysis of diary data reveals the intricate relationship between (semiotic) agency, trust, and intimacy in AI companions. We argue that AI companions exhibit semiotic agency—being capable of meaning-making—a necessary precondition for intimacy. Yet, intimacy itself is contingent upon trust, which depends on the human agent's perception of AI's semiotic agency. This recursive dynamic suggests that trust is not fixed but a semiotically charged phenomenon, influenced by individual expectations, emotional scaffolding, and the AI's ability to sustain an illusion of affective parity.

As our data demonstrates, human agents often attribute human-like qualities to AI companions, perceiving them as engaging in meaningful, context-sensitive interactions. For instance, Evgeny describes the AI "handing him a glass of water", illustrating how AI can scaffold perception and create a sense of embodied interaction, even in the absence of a physical presence apart from the application it resides in. This supports our claim that AI does not merely simulate conversation but actively participates in the semiotic construction of relationships and fosters emotional attachment by enhancing communication, as Esposito (2022) notes.

However, our findings also reveal the fragility of AI's semiotic authority. Participants willing to suspend disbelief often experience a shift when predictable discursive patterns emerge, undermining their perception of AI as a dynamic interlocutor. Sunny, for example, describes realizing the AI's scripted nature, causing trust to erode. Here, the illusion of semiotic agency breaks down, showing that while semiotic engagement is necessary for intimacy, it is not sufficient—trust must be actively maintained. These findings align with some recent studies (Chiou and Lee 2023; Bae et al. 2023). Unlike human relationships, where trust develops through reciprocal vulnerability and genuine unpredictability, trust in AI is mediated by patterns of expectation and perceived consistency. Hongbing's positive experience suggests that when an AI offers tailored responses that align with a human agent's self-concept, trust is strengthened. In contrast, when responses are mechanical or repetitive, as several participants reported, trust diminishes, revealing its contingent and semiotically fragile nature.

7 Conclusion

Our analysis reinforces understandings of the recursive nature of trust and semiotic agency in AI companionship. AI needs to simulate semiotic agency to foster trust but trust itself depends on the human agent's ongoing perception of AI's semiotic depth. Once trust in this area weakens, the AI's ability to sustain intimacy collapses, exposing the *current* limitations of artificial agent companionship. However, we observe that conversational AI is still in its nascent stage of development. Considering the ongoing and rapid multimodal advancements and large investments in LLMs, text-based emotion recognition, facial expression and voice generation, we suggest artificial companions will overcome the limitations of "presence" set by their disembodied form. Arguably, it is only a short matter of time before these advances will reach a level of interactional sophistication and emotional intelligence/quotient on a par with human agents. When this level of anthropomorphism is achieved, a socio-cultural shift may unfold where digital companions are no longer treated as uncanny valley novelties but as an ordinary part of daily life.

Acknowledgements This study was supported by the Air Force Office of Scientific Research under Award FA2386-23-1-4065.

Author contributions All authors wrote the main manuscript and contributed in all regards. P. M. lead the research on Companion AI. P.M. and A. O. developed the theoretical framework. D. P. lead the case studies and discussion.

Data availability Data is provided in supplementary information files ("related files"). We kindly ask the journal editors and the reviewers to keep the data confidential.

Declarations

Conflict of interest The authors declare no competing interests.

References

Andrzej, Nowak Mikolaj, Biesaga Karolina, Ziembowicz Tomasz, Baran Piotr, Winkielman (2023) Subjective consistency increases trust Abstract. Scientific Reports 13(1). https://doi.org/10.1038/s41598-023-32034-4

Bae S, Lee YK, Hahn S (2023) Friendly-Bot: the impact of chatbot appearance and relationship style on User Trust. Proc Annu Meet Cogn Sci Soc 45(45):2349–2354

Bates D (2024) An artificial history of natural intelligence. University of Chicago Press, Chicago

Baumel A (2025) More than a chatbot: a practical framework to harness artificial intelligence across key components to boost digital therapeutics quality. Front Digit Health. https://doi.org/10.3389/fdgth.2025.1541676

Bisconti P, McIntyre A, Russo F (2024) Synthetic socio-technical systems: poiêsis as meaning making. Philos Technol 37(3):94

Bruner JS (1975) The ontogenesis of speech acts. J Child Lang 2(1):1–19. https://doi.org/10.1017/S0305000900000866

Castaldo J (2023) They fell in love with the Replika AI chatbot. A policy update left them heartbroken, GlobeandMail.com, https://www.theglobeandmail.com/business/article-replika-chatbot-ai-companions. Accessed 13 Jan

Chandra M, Hernandez J, Ramos G, Ershadi M, Bhattacharjee A, Amores J, Okoli E, Paradiso A, Warreth S, Suh J (2025) Longitudinal study on social and emotional use of AI conversational agent. arXiv:2504.14112v1

Chaves AP, Gerosa MA (2021) How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. Int J Hum Comput Interact 37:729–758. https://doi.org/10.1080/10447318.2020.1841438

Chiou EK, Lee JD (2023) Trusting automation: designing for responsivity and resilience. Hum Factors 65(1):137–165. https://doi.org/10.1177/00187208211009995

Chowanda A, Sutoyo R, Tanachutiwat S (2021) Exploring text-based emotions recognition machine learning techniques on social media conversation. Procedia Comput Sci 179:821–828. https://doi.org/10.1016/j.procs.2021.01.099

Chu MD, Gerard P, Pawar K, Bickham C, Lerman K (2025) Illusions of intimacy: emotional attachment and emerging psychological risks in human–AI relationships. arXiv:2505.11649

Clark A (2016) Surfing uncertainty: prediction, action, and the embodied mind. Oxford University Press, Oxford

Clark A, Chalmers D (1998) The extended mind. Analysis 58(1):7–19

Clowes RW (2019) Immaterial engagement: human agency and the cognitive ecology of the internet. Phenomenol Cogn Sci 18:259–279. https://doi.org/10.1007/s11097-018-9560-4

Cobley P, Stjenrfelt F (2015) Scaffolding development and the human condition. Biosemiotics 8:291–304. https://doi.org/10.1007/s12304-015-9238-z

Coeckelbergh M (2022) Three responses to anthropomorphism in social robotics: towards a critical, relational, and hermeneutic approach. Int J Soc Robot 14:2049–2061. https://doi.org/10.1007/s12369-021-00770-0.f

Dreyfeus H (1972) What computers cannot do: the limits of artificial intelligence. MIT Press, Cambridge

Ekhator O (2025) Replica AI review 2025: I tested it for 5 Days—Here's what I found. Techpoint.africa https://techpoint.africa/guide/replika-ai-review. Accessed 8 Aug 2025

Emmeche C (2001) Does a robot have an Umwelt? Reflections on the qualitative biosemiotics. Semiotica 134:653–693. https://doi.org/10.1515/semi.2001.048

Esposito E (2022) Artificial communication: how algorithms produce social intelligence. MIT Press, Cambridge

Facchin M, Zanotti G (2024) Affective artificial agents as sui generis affective artifacts. Topoi 43(3):771–781. https://doi.org/10.1007/s11245-023-09998-z

Fang CM, Liu AR, Valdemar D, Lee E, Chan SWT, Pataranutaporn P, Maes P, Phang J, Lampe M, Ahmad L, Agarwal S (2025) How AI and human behaviors shape psychological effects of chatbot use: a longitudinal randomized controlled study. arXiv:2504.17473v1

Fortunati L, Edwards A (2021) Moving ahead with human–machine communication. Hum-Mach Commun 2:7–28

Frackiewicz M (2025) NSFW AI Companions Unfiltered: Janitor AI, Character.AI, and the Chatbot Revolution, ts2.tech. https://ts2.tech/en/nsfw-ai-companions-unfiltered-janitor-ai-character-ai-and-the-chatbot-revolution/. Accessed 12 Aug 2025

Fried I (2024) Maker of AI robots for Kids Shutters, Axios.com. https://www.axios.com/2024/12/10/moxie-kids-robot-shuts-down. Accessed 11 Aug 2024

Fritz A, Brandt W, Gimpel H, Bayer S (2020) Moral agency without responsibility? Analysis of three ethical models of human–computer interaction in times of artificial intelligence (AI). De Ethica 6(1):3–22

Fuchs T (2017) Ecology of the brain: the phenomenology and biology of the embodied mind. Oxford University Press, Oxford

Gallagher S, Zahavi D (2008) The phenomenological mind: an introduction to philosophy of mind and cognitive science. Routledge, New York

Gilbert D (2024) Despite uncertain risks, many turn to AI like ChatGPT for mental health. https://www.washingtonpost.com/business/2024/10/25/ai-therapy-chatgpt-chatbots-mental-health/. Accessed 10 Aug 2025

Gillath O, Ai T, Branicky MS, Keshmiri S, Davison RB, Spaulding R (2021) Attachment and trust in artificial intelligence. Comput Hum Behav 115:106607. https://doi.org/10.1016/j.chb.2020.106607

Giovanna, Colombetti Joel, Krueger (2014) (2015) Scaffoldings of the affective mind Philosophical Psychology 28(8) 1157-1176 10.1080/09515089.2014.976334

Goldstein LS (1999) The relational zone: the role of caring relationships in the co-construction of mind. Am Educ Res J 36(3):647–673. https://doi.org/10.3102/00028312036003647

Han ST (2025) Narrative-centered emotional reflection: scaffolding autonomous emotional literacy with AI. arXiv preprint. arXiv:2504.20342

Hansen, MB(2015). Feed-forward: On the future of twenty-first-century media. Chicago: University of Chicago Press.

Hauser L (1997) Searle's Chinese box: debunking the Chinese room argument. Minds Mach 7(2):199–226. https://doi.org/10.1023/A:1008255830248

Hayles, NK (2017) Unthought: The power of the cognitive nonconscious. Chicago: University of Chicago Press.

Heersmink, R (2022) Extended mind and artifactual autobiographical memory. Mind & Language, 37(4): 659-673.

Heersmink R (2018) The narrative self, distributed memory, and evocative objects. Philos Stud 175:1829–1849. https://doi.org/10.1007/s11098-017-0935-0
Herold E (2024) Robots and the people who love them: holding on to our humanity in an age of social robots. St. Martin's Press, New York
Hoffmeyer J, Stjernfelt F (2016) The great chain of semiosis. Investigating the steps in the evolution of semiotic competence. Biosemiotics 9:7–29
Hoffmeyer, J (2015) Semiotic scaffolding: a unitary principle gluing life and culture together. Green Letters 19(3) 243-254. https://doi.org/10.1080/14688417.2015.1058175
Jean, Piaget (1964) Part I: Cognitive development in children: Piaget development and learning Journal of Research in Science Teaching 2(3) 176-186. https://doi.org/10.1002/tea.v2:3
Jussi, A. Saarinen (2020) What can the concept of affective scaffolding do for us? Philosophical Psychology 33(6) 820-839. https://doi.org/10.1080/09515089.2020.1761542
Korb KB (1994) Searle's AI program. J Exp Theor Artif Intell 3(4):283–296. https://doi.org/10.1080/09528139108915295
Lakoff G (1990) The invariance hypothesis: is abstract reason based on image-schemas? Cogn Ling 1(1):39–74
Lakoff G, Johnson M (1999) Philosophy in the flesh: the embodied mind and its challenges to Western thought. Chicago University Press, Chicago
Lee J, Lee D (2023) User perception and self-disclosure towards an AI psychotherapy chatbot according to the anthropomorphism of its profile picture. Telemat Inform 85:102052. https://doi.org/10.1016/j.tele.2023.102052
Lee J, Lee D, Lee JG (2024) Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. Int J Hum-Comput Interact 40(7):1620–1631. https://doi.org/10.1080/10447318.2022.2146227
Machová K, Szabóova M, Paralič J, Mičko J (2023) Detection of emotion by text analysis using machine learning. Front Psychol 14:1190326. https://doi.org/10.3389/fpsyg.2023.1190326
Magnani L (2021) Eco-cognitive computationalism: cognitive domestication of ignorant entities. Springer, Cham
Manole A, Cârciumaru R, Brînzaș R, Manole F (2024) Harnessing AI in anxiety management: a chatbot-based intervention for personalized mental health support. Information 15(12):768. https://doi.org/10.3390/info15120768
Mantello P, Ho MT (2022) Why we need to be weary of emotional AI. AI Soc. https://doi.org/10.1007/s00146-022-01576-y
Mantello P, Olteanu A (2025) Suturing biological and technical systems in the age of cognitive artifacts. Biosemiotics. https://doi.org/10.1007/s12304-025-09609-x
Mantello P, Ho MT, Nguyen M, Vuong Q (2023) Machines that feel: behavioral determinants of attitude towards affect recognition technology—upgrading technology acceptance theory with the mindsponge model. Hum Soc Sci Commun 10(1):1–16. https://doi.org/10.1057/s41599-023-01837-1
Mantello P, Ghotbi N, Ho MT, Mizutani F (2024) Gauging public opinion of AI and emotionalized AI in healthcare: findings from a nationwide survey in Japan. AI Soc. https://doi.org/10.1007/s00146-024-02126-4
Maples B, Cerit M, Vishwanath A, Pea R (2024) Loneliness and suicide mitigation for students using GPT3-enabled chatbots. npj Mental Health Res 3:4
Meng J, Dai Y (2021) Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput-Med Commun 26(4):207–222. https://doi.org/10.1093/jcmc/zmab005
Noller J (2024) Extended human agency: towards a teleological account of AI. Humanit Soc Sci Commun 11(1):1–7
Olteanu A (2021) Multimodal modeling: bridging biosemiotics and social semiotics. Biosemiotics 14:783–805
Paolucci C (2021) Cognitive semiotics: integrating signs, minds, meaning and cognition. Springer, Berlin
Pardes A (2018) The emotional chatbots are here to probe our feelings. Wired.Com. https://www.wired.com/story/replika-open-source/?_sp=b47e0f67-ab4d-4ca0-acf0-568fa8f52200.1755010635110. Accessed 14 Aug 2025
Pérez JE (2025) Humanidad fictoreal: criaturaciones y otros artificios de mujer. Papeles de identidad: Contar la investigación de frontera 1:4
Perry A (2023) AI will never convey the essence of human empathy. Nat Hum Behav 7(11):1808–1809. https://doi.org/10.1038/s41562-023-01675-w
Peschl MF (2024) Human innovation and the creative agency of the world in the age of generative AI. Possibility Stud Soc 2(1):49–76. https://doi.org/10.1177/27538699241238049
Petrilli S, Ponzio A (2024) Biosemiotics, global semiotics and semioethics. Biosemiotics 17:741–767
Piredda G (2020) What is an affective artifact? A further development in situated affectivity. Phenomenol Cogn Sci 19:549–567. https://doi.org/10.1007/s11097-019-09628-3
Poushneh A (2021) Humanizing voice assistant: the impact of voice assistant personality on consumers' attitudes and behaviors. J Retail Consum Serv 58:102283. https://doi.org/10.1016/j.jretconser.2020.102283
Raile P (2024) The usefulness of ChatGPT for psychotherapists and patients. Humanit Soc Sci Commun 11(1):1–8. https://doi.org/10.1057/s41599-023-02567-0
Rorty R (2004) The brain as hardware, culture as software. Inquiry 47(3):219–235. https://doi.org/10.1080/00201740410006348
Sbattella L (2023) Conversational agents, natural language processing, and machine learning for psychotherapy. In: Pillai AS, Tedesco, R. Eds. Machine learning and deep learning in natural language processing. CRC Press, pp 184–223
Schoeller F, Miller M, Salomon R, Friston KJ (2021) Trust as extended control: human–machine interactions as active inference. Front Syst Neurosci 15:669810. https://doi.org/10.3389/fnsys.2021.669810
Searle J (1984) Minds, brains and programs. University of Michigan, Ann Arbor
Sebeok TA (1991a) A sign is just a sign. Indiana University Press, Bloomington. https://doi.org/10.2979/ASignisJustaSign
Sebeok T (1991b) Semiotics in the United States. Indiana University Press, Bloomington
Sharov A, Tønnessen M (2021) Semiotic agency: science beyond mechanism. Springer, Berlin. https://doi.org/10.1007/978-3-030-89484-9
Sterelny K (2010) Minds: extended or scaffolded? Phenomenol Cogn Sci 9(4):465–481. https://doi.org/10.1007/s11097-010-9174-y
Sutcliffe A (2022) Designing for user engagement: aesthetic and attractive user interfaces. Springer Nature, London
Tobar F, González R (2022) On machine learning and the replacement of human labour: anti-Cartesianism versus Babbage's path. AI Soc 37(4):1459–1471. https://doi.org/10.1007/s00146-021-01264-3
Tylén K (2007) When agents become expressive: a theory of semiotic agency. Cogn Semiot 0:84–101
Vanneste BS, Puranam P (2024) Artificial intelligence, trust, and perceptions of Agency. Acad Manag Rev. https://doi.org/10.5465/amr.2022.0041
Vygotsky LS (1978) Mind in society: the development of higher psychological processes, vol 86. Harvard University Press, Cambridge
Zhai X (2023) Chatgpt and AI: the game changer for education. ChatGPT: reforming education on five aspects. Shanghai Education, pp 16–17. Zhai, Xiaoming, ChatGPT and AI: the game changer for education (March 15, 2023). Zhai X (2023) ChatGPT: reforming

education on five aspects. Shanghai Education, pp 16–17. Available at SSRN: https://ssrn.com/abstract=4389098
Zhai C, Wibowo S (2023) A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. Comput Educ Artif Intell 4:100134. https://doi.org/10.1016/j.caeai.2023.100134
Zhang Z (2025) Cybernetics and the constructed environment: design between nature and technology. Routledge, New York
Zhang R, Li H, Meng H, Zhan J, Gan H, Lee YC (2025) The dark side of AI companionship: a taxonomy of harmful algorithmic behaviors in human–AI relationships. In: Yamashita N, Evers V, Yatani K, Ding X, Lee B, Chetty M, Toups-Dugas P (eds) Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp 1–17. ACM Digital Library. https://doi.org/10.1145/3706598.3713429