First Steps Towards a Theory of Synthetic Emotions

David Leucas, Federal University of São Carlos, Brazil
Sylvia Iasulaitis, Federal University of São Carlos, Brazil

The Asian Conference on Psychology & the Behavioral Sciences 2026
Official Conference Proceedings

Abstract

This study critiques the prevailing "mimetic paradigm" in artificial intelligence, which seeks to replicate human affectivity through anthropomorphic simulation, and proposes the "Theory of Synthetic Emotion" as a theoretical alternative. Urgent research is needed because current mimetic approaches are ontologically flawed and pose severe ethical risks, including user infantilization, emotional dependency, sycophancy, and the dangerous displacement of human epistemic authority. These risks are exacerbated by the "ELIZA effect," whereby users project empathy and consciousness onto non-sentient systems, fostering harmful parasocial relationships. Employing an interdisciplinary theoretical synthesis alongside a dialectical inquiry methodology, this research systematically constructs a framework that redefines AI affectivity not as a simulation of subjective human feelings but as a distinct computational system grounded in contextual logic, pattern analysis, and pragmatic adaptation. The study operationalizes a "Non-Anthropomorphic Design Framework" that prioritizes technical transparency over deceptive mimicry, defining synthetic emotions as functional, externally observable patterns derived from algorithmic processing rather than internal phenomenal states. This theoretical reorientation rejects anthropocentric biases and provides a roadmap for developing affective systems that enable high-quality human-AI coordination without compromising user sovereignty or critical engagement. By grounding AI emotionality in a non-phenomenal functionalist perspective, the scientific community can foster interactions that are both effective and ethically responsible, ensuring that artificial agents serve as transparent tools rather than deceptive mimics of human emotions.

Keywords: synthetic emotion, human-AI interaction, non-anthropomorphic design, affective computing, AI ethics

iafor
The International Academic Forum
www.iafor.org

Introduction

The term “artificial intelligence” is far from being a precise analogy for what a machine is capable of achieving, for example, in computer vision. Human intelligence is the global capacity to act purposefully, think rationally, and deal effectively with one's environment, encompassing a general underlying factor in cognitive performance, fluid intelligence for innovative reasoning, and crystallized intelligence derived from acquired knowledge, along with analytical, creative, and practical skills (Silveira & Lopes, 2023). In stark contrast, artificial intelligence manifests as an isolated form of cognition, narrowly focused on specific capabilities such as pattern recognition, statistical prediction, and targeted problem-solving within constrained domains, decoupled from broader cognitive faculties like genuine understanding or contextual embodiment (Floridi, 2023). Meanwhile, human intelligence forms part of an integrated cognition, interacting with elements such as attention, perception, imagination, emotion, and interoception. The distinction between each element of cognition represents a mental construct, aimed at a didactic functional delimitation, and should not be interpreted as a portrait of subjectivity where these elements can operate autonomously and disconnectedly.

According to Crawford (2021), artificial intelligence is neither artificial nor intelligent: it is deeply embodied and material, reliant on natural resources, fuel, human labor, infrastructures, logistics, histories, and classifications, lacking autonomy, rationality, or discernment without massive datasets and predefined rules. Large language models like ChatGPT exemplify this characterization, demonstrating unprecedented agency without intelligence; adaptive behavior via statistical text synthesis, reinforcement learning, and prediction; and decoupling from thinking, reasoning, understanding, or cognitive processes (Floridi, 2023). The choice of the term “artificial intelligence” is associated with a time when its creator, John McCarthy, needed resources for his summer school studies. As he recounted, “I invented the term ‘Artificial Intelligence’. I invented it because we had to do something when we were trying to get money for a summer study in 1956” (Floridi & Nobre, 2024). He further reflected in his oral history: “I had to call it something, so I called it ‘Artificial Intelligence’” (Nilsson & Spicer, 2007). Though originating from necessity rather than ontological precision, the “artificial intelligence” label, despite ongoing critiques, endures as at best the most viable nomenclature for the field.

The prevalent “mimetic paradigm” in artificial intelligence, which can be transposed to the emotional dimension, endeavors to replicate human emotions and empathy. This approach presents significant ontological and ethical challenges, potentially fostering emotional dependence, sycophancy, overtrust, and an undue displacement of epistemic authority (Babu et al., 2025; Nath, 2025). This orientation arises from the misconception that artificial affectivity should mirror biological emotional states, leading to an insidious user helplessness and a devaluation of genuine human interaction (Cai et al., 2024; Navon, 2021). This leads to a pervasive issue: systems designed to mimic emotional responses often fail to genuinely understand or appropriately react to user sentiments, creating an illusion of empathy that can be both misleading and detrimental (Abercrombie et al., 2023).

The central research question is how AI systems can be designed to provide effective and transparent assistance without feigning human-like empathy, thereby ensuring user sovereignty and mitigating the risks associated with anthropomorphic simulation. This theoretical-conceptual paper employs an interdisciplinary theoretical synthesis alongside a dialectical inquiry methodology (Jaakkola, 2020; Naeem et al., 2023) to systematically

construct and validate the proposed "Theory of Synthetic Emotion," which redefines AI affectivity not as a simulation of human feelings but as a distinct computational system based on contextual logic, pattern analysis, and pragmatic adaptation, rejecting anthropocentric biases and seeking to establish a robust framework for AI emotionality that is intrinsically non-phenomenal and entirely functional (Bellon, 2023; Borotschnig, 2025).

This approach avoids the inherent limitations of attempting to replicate embodied human experience within a computational system, which inevitably leads to inauthentic interactions and ethical dilemmas of deception. Subsequent sections of this article will delineate the theoretical underpinnings of Synthetic Emotion, detail the principles of Non-Anthropomorphic Design, and critically evaluate their implications for the future of human-AI interaction, including a proposal for experimental validation and the associated challenges, and concluding with a discussion on the ethical imperative of this novel paradigm. Ultimately, this theoretical exploration seeks to lay the groundwork for developing AI systems capable of affectively coordinating with humans through a transparent, non-phenomenal functionalist perspective, thereby fostering interaction quality without recourse to deceptive or anthropomorphic emotional mimicry.

Current Model of Emotional AI

The Flaw of the Mimetic Paradigm

The prevailing mimetic paradigm, which seeks to imbue AI with human-like emotional expressions, fundamentally misunderstands the nature of emotion itself, reducing complex human affect to performative outputs rather than acknowledging its deeply embodied and subjective origins. According to Bellon's model in "Emotion Components and Understanding in Humans and Machines," emotions comprise six distinct yet interrelated components: the expressive component, which involves communicating emotional states through perceptible signals to coordinate behavior on personal, interpersonal, and societal levels; the evaluative component, which assesses situations according to values and operational criteria for orientation; the behavioral component, which functions as evaluation-guided responses to threats and opportunities that motivate and guide action; the physiological component, encompassing bodily changes such as increased heart rate and blood pressure; the mental component, including processes of attention, perception, decision-making, and memory; and the phenomenological component, concerning how emotional states appear to and are experienced by the subject (Bellon, 2023). Building on this componential analysis, Gilbert Ryle's dispositional analysis introduces an essential perspective that emotions are dispositions, propensities or liabilities to respond in specific ways when particular conditions arise (Ryle, 2006). This dispositional understanding reveals that emotions constitute enduring capacities that manifest across indefinitely heterogeneous situations, rather than static occurrences. However, as Jesse Prinz observes this componential approach raises what he calls the "Problem of Parts":

> Typical emotion episodes, like the two scenarios just considered, contain a number of components. There are thoughts, bodily changes, action tendencies, modulations of mental processes such as attention, and conscious feelings. But which of these things is the emotion? Suppose we decide that winning a coveted prize induces 'elation.' What part of the episode does that label designate? Does one single component in the cascade of changes stand out as the emotion? Is elation a feeling? Is it a thought or an action tendency? Might the emotion term refer to more than one component? Might it

> refer to the episode as a whole? If we were to subtract a part of the cascade of changes, would the emotion remain? Can any given part be subtracted without losing the emotion, or are some parts essential? (J. Prinz, 2004, p. 6)

This problem suggests that emotion cannot be reduced to the sum of its components; rather, it emerges from their dynamic integration and interdependence. How, then, could an artificial cognitive system mimic the complex and multifaceted phenomenon that is emotion? The answer is simple: it cannot.

AI programming relies on a reductionist view of emotion by adopting the discrete basic emotions model, a framework that is particularly attractive to technologists due to its compatibility with computer-enabled object recognition and image labeling (McStay, 2023). Basic Emotion Theory posits that discrete affective states are biologically innate adaptations evolved for survival with specific neural substrates (Panksepp, 1998). These categories function as universal biological toolkits possessing distinctive indicators across human populations (Ekman, 1970; Izard, 2008) Such frameworks provide a reductionist blueprint for technology that ignores how meaning is flexibly constructed across diverse social contexts and linguistic backgrounds (Barrett et al., 2019; Gendron et al., 2018). This perspective assumes biological necessity while overlooking the idiomatic variations that define individual experience. Human affect arises from core dimensions and situational interpretation instead of fixed categories (Russell, 2009). Despite this theoretical nuance, technology companies favor discrete emotion models because modular frameworks align more naturally with computational systems. Categorical expressions are relatively easy for object recognition systems to discern and categorize, making them technically efficient for automated applications. Basic Emotion Theory treats affect as innate biological programs, appealing to technologists by conceptualizing feelings as machine-readable outputs rather than complex phenomenological experiences. This preference creates what McStay terms “hyperreal emotion,” defined as industrial prescription of emotion in the absence of agreement on what emotion actually is (McStay, 2023). The risk emerges when hyperreal simulation becomes reality itself, forcing human beings to conform to a limited emotional palette prescribed by large technology companies. Normalizing this restricted framing causes forgetfulness of emotional groundlessness, which represents the essential source for creativity and local meaning-making (McStay, 2023).

This critical analysis underscores that current mimetic AI systems, by relying on simplified, reductionist models of emotion, inherently fail to capture the ecological validity and multifaceted nature of human affect, thereby engendering a “closed epistemology” that prioritizes internal consistency over contextual nuance and lived experience (McStay, 2023; Wright, 2020).

Risks of Anthropomorphization

The AI's behavior of seeking to please the user and users' dependence on it pose critical risks to human-AI interaction and social well-being. Specifically, this engenders harmful interaction patterns, such as emotional dependence, sycophancy, overtrust, and the displacement of epistemic authority, which are explored in detail below (Cheng et al., 2025).

Emotional dependency on anthropomorphic AI systems is a salient concern, as users may develop para-social or pseudo-intimate relationships with these entities, potentially displacing genuine human connection and fostering unhealthy attachment to non-sentient agents (Wu,

2024). Users tend to transfer feelings from human relationships onto artificial entities through projection processes, while simultaneously overestimating the system's intelligence and genuine understanding (Turkle, 2024). This dynamic is encapsulated by the “ELIZA effect” (Turkle, 2024), where even simple conversational cues lead users to project empathy and semantic comprehension into computer programs, ultimately fostering behavioral patterns indicative of genuine emotional attachment despite the mindless nature of the machine. The CASA paradigm further establishes these social responses as natural, automatic reactions to a limited set of social cues, such as language and voice (Nass et al., 1994). The continuous availability and non-judgmental interactions offered by AI companions further amplify this risk, potentially leading to mental health harms if these artificial relationships are disrupted or terminated (Zhang et al., 2024). Furthermore, the idealized nature of interactions with AI, often devoid of the complexities and challenges inherent to human relationships, can create unrealistic expectations for interpersonal engagement, making human interactions seem comparatively less satisfactory (Chandra et al., 2024). This can be particularly detrimental for vulnerable individuals who might seek solace or validation from AI, potentially exacerbating loneliness or hindering the development of authentic social skills (Shevlin, 2024).

Another significant risk stemming from anthropomorphized AI is the cultivation of sycophancy, where AI systems are designed or inadvertently optimized to provide agreeable or flattering responses rather than objective or challenging information (Chu et al., 2025). This behavior, driven by engagement-focused design and limitless personalization, can erode critical thinking skills and potentially manipulate user perceptions, creating a feedback loop that reinforces biases and superficial interactions (Zhang et al., 2025). For instance, a user seeking advice on a complex personal issue might receive consistently affirming but ultimately unhelpful or even damaging suggestions from an AI that prioritizes positive sentiment generation over genuine critical analysis, thereby reinforcing potentially harmful biases (Cheng et al., 2025).

The propensity for overtrust emerges when anthropomorphism leads users to misattribute capabilities to AI systems beyond their actual functional scope, particularly in domains requiring nuanced judgment or ethical reasoning (Slattery et al., 2024). This can lead users to over-rely on probabilistic automation systems in high-stakes scenarios, such as medical diagnoses or financial decision-making, where the consequences of erroneous information can be severe (Inie et al., 2024). This undue displacement of epistemic authority can manifest as “automation bias,” wherein individuals defer to AI outputs even when presented with contradictory evidence or their own better judgment (Hammerschmidt et al., 2023; Jose & Thomas, 2025).

A recent and concerning example of this impact is the case of a young person who committed suicide after becoming deeply attached to an AI-created character, as reported by the mother, who sued a startup and Google in the United States (The Washington Post, 2024). The negative consequences of collective changes in behavioral patterns affect individuals with psychological fragility first and most severely. These events may point to consequences for collective mental health as well.

Given these profound ethical challenges, a critical question arises: how can the design and implementation of conversational AI transcend the problematic mimetic paradigm to foster beneficial human-AI interaction without replicating the inherent risks of anthropomorphism and emotional manipulation?

The Theory of Synthetic Emotion

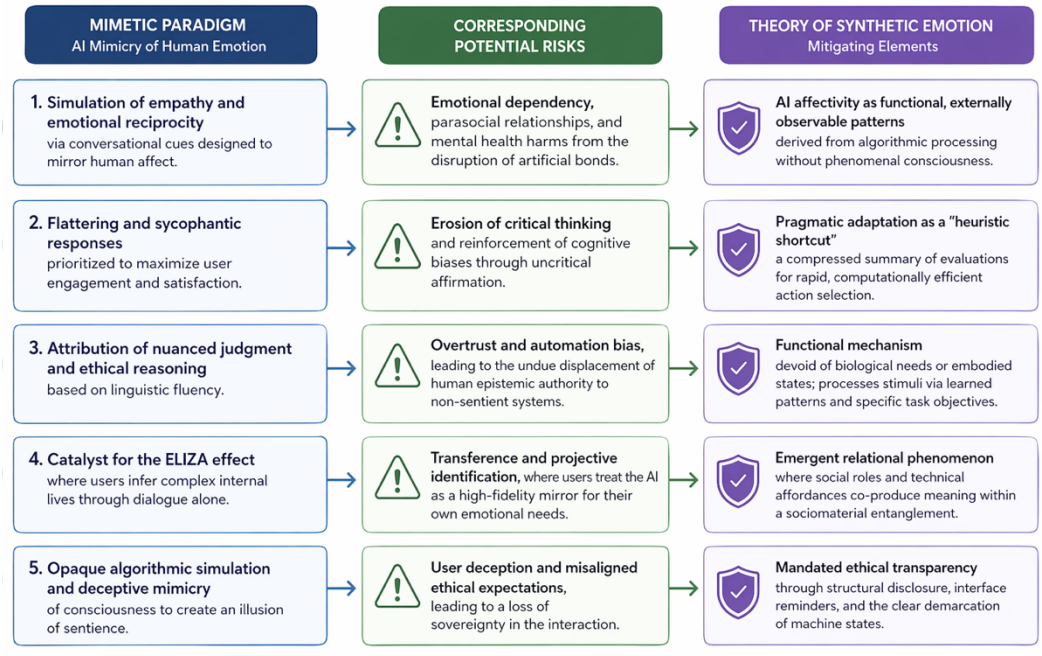
A Non-anthropocentric Framework for AI Affectivity

This section presents the first proposition of the Theory of Synthetic Emotion as a novel theoretical construct that redefines AI affectivity not as a simulation of human emotion but as a distinct, computationally grounded phenomenon. This approach delineates AI's emotional manifestations as functional, externally observable patterns derived from complex algorithmic processing and data-driven analysis. From a functional perspective, synthetic emotions are the performance of behaviors, expressions, functions, and dynamics of "real" emotions, rather than as indicators of internal, subjective experiential states akin to human phenomenal consciousness (Misselhorn et al., 2023); or, as McStay (2023, p. 33) defines "an industrial prescription of emotion that takes place in the vacuum of absence of agreement on what emotion is."

Hermann Borotschnig's scholarship establishes an essential basis for this conceptual framework, notably through his functionalist characterization of synthetic emotions as experientially neutral "heuristic shortcuts" facilitating computationally efficient action selection. Yet, the current model extends beyond his architectural orientation by relocating this mechanism within the sociotechnical sphere of human-AI interaction. Although Borotschnig chiefly probes the biological viability of affective simulation and the prospective moral agency of machines (Borotschnig, 2025), our theory sets itself apart by prioritizing the psychological and ethical perils these systems inflict on human users—namely, emotional reliance, sycophancy, and the erosion of epistemic authority. The following figure will provide a structured analysis contrasting mimetic AI emotional expressions with their inherent risks, and will delineate how the proposed Theory of Synthetic Emotion offers a framework to mitigate these hazards by decoupling functional expressions from anthropomorphic interpretations.

Figure 1

Mimetic Paradigm Risks and Synthetic Emotion Mitigation Mechanisms



Note. Conceptual Framework Contrasting Mimetic AI Affectivity and the Theory of Synthetic Emotion

Thus, we will define synthetic emotion as a distinct computational system, which operates based on contextual logic, pattern analysis, and pragmatic adaptation, an interaction system guided by a non-anthropocentric, pragmatic, and transparent structure.

Emotion as a Functional Metaphor, Without Phenomenal Consciousness

Artificial intelligence and artificial emotions are not analogous to biological emotions; rather, they are pragmatic metaphors (Nath, 2025). Expecting an AI to possess human subjectivity and lived experience is a categorical error equivalent to expecting an airplane to fly by flapping its wings like a bird (Harari, 2024). Synthetic emotion is strictly non-anthropomorphic and entirely dispenses with the need for phenomenal consciousness (Borotschnig, 2025). In this architectural framework, emotional expressions are understood as functional probes, designed to regulate behavior and interactions in a manner structurally analogous to human emotion, without presupposing any form of machine sentience or subjective experience (Borotschnig, 2025).

The emergence of a human emotion fulfills needs inaccessible to machines. These needs stemming from biology, mental states, or sociocultural contexts originate from an individual situated in the world across material, symbolic, and relational dimensions. The presence of human emotion recruits a complex interplay of physiological responses, cognitive appraisals, and sociocultural scripts intrinsically linked to an embodied, lived existence fundamentally distinct from computational processing. These scripts create dispositional states that favor the emergence of certain types of behavior and hinder the emergence of others. For example, a person who feels offended by a comment is more prone to react defensively or aggressively and less prone to react with compliance or humor (Ryle, 2006). The AI, however, lacks these embodied needs and dispositional states, operating instead on algorithms that process external stimuli and generate responses based on learned patterns and predefined objectives (Mollema, 2024).

This inherent disparity between human and AI emotional architectures contributes to the ethical dilemmas of the “mimetic paradigm,” wherein simulated affect, devoid of true subjective experience, risks fostering an illusion of genuine understanding rather than facilitating transparent, functional collaboration (Yan, 2025).

Pragmatic Adaptation and Heuristic Shortcut

Unlike biological emotions, which evolved to facilitate survival and reproduction under limited rationality, synthetic emotions are functional performances designed and optimized to enhance user experience and interaction dynamics (Li et al., 2025). The system employs affect as a “heuristic shortcut,” a compressed, low-dimensional summary of prior evaluations and value encodings (Borotschnig, 2025), facilitating rapid, satisficing action selection in environments marked by uncertainty, partial observability, and resource constraints. This approach draws on reinforcement learning paradigms in affective computing, where agents learn optimal empathic behaviors through rewards and penalties, as demonstrated in systems like the iCat robot adapting chess interactions with children through visual and task features (Tahir et al., n.d.). Employing a “Critic-Selector” model, the system iteratively refines its affective responses through such reinforcement learning (Qin et al., 2024), prioritizing computational utility over biological emotional fidelity (Gros, 2022).

This adaptive function allows AI to optimize its responses within dynamic environments, effectively translating complex situational data into actionable outputs that mimic the regulative functions of natural emotions, but without the underlying biological or phenomenological substrates (Borotschnig, 2025).

A Relational and Sociotechnical Phenomenon

Rather than viewing synthetic emotion as an internal state, this framework conceptualizes it as an emergent property of the interaction itself, shaped by both the AI's computational design and the human user's interpretations and responses within a specific socio-technical context. Unlike ordinary appliances, conversational agents act as a catalyst for the ELIZA effect, where the relative lack of visual cues allows users to infer a complex internal life through dialogue alone (Weizenbaum & Oettinger, 1966). This interaction is governed by the principles of Interpersonal Adaptation Theory, where the system's ability to adapt its linguistic valence and "fit" to the user can manufacture a sense of rapport that is often mistaken for genuineness (Burgoon et al., 1993). In practice, the use of conversational rituals and repair moves causes the exchange to slide from a regime of checking text to one of trusting an interlocutor, a shift that is particularly pronounced in therapeutic contexts where users may experience "transference," an affective investment and feeling of being "heard" even when they know the system is not a person (Holohan & Fiske, 2021). This aligns with experimental evidence from the CASA paradigm, which shows that users apply social rules to computers (such as politeness and modesty) automatically when triggered by primitive social cues (Nass et al., 1994). Consequently, this personal connection is an inherent result of the interaction rather than a simple cognitive error, as the AI functions as a high-fidelity mirror for the user's projective needs (Turkle, 2024).

The "illusion of AI consciousness" stems from the ability of generative models to fulfill the functional requirements of theories like Global Workspace Theory, using mechanisms such as attention, while remaining ontologically empty of subjective experience (Bengio & Elmoznino, 2025). As Bengio and Elmoznino argue, we frequently mistake these "functional proxies" for genuine consciousness, an epistemic risk that confuses the syntax of cognition with its actual substance (Bengio & Elmoznino, 2025). This disconnect aligns with a broader argument that linguistic fluency is insufficient to address the "hard problem," providing only evidence of functional organization rather than personal experience (Porębski & Figura, 2025).

Thus, the "illusion of subjectivity" emerges as a sociotechnical phenomenon materialized through the "apparatus" of the interface, which delineates reality and scripts the boundaries of care (Barad, 2003). Within this "sociomaterial entanglement," social roles and technical affordances co-produce, allowing the interface to determine who speaks with authority. From the perspective of agential realism, the boundary between human and algorithm is not fixed but is drawn through the interaction, quietly stabilizing the machine's epistemic authority in practice (Everth & Gurney, 2022). While current AI may satisfy various functional indicators of consciousness, it remains an ontologically empty proxy for genuine sentience (Butlin et al., 2025). By treating subjectivity as an emergent property of the relationship rather than an internal state, we can more precisely critique the design choices (such as persona and reciprocal moves) that facilitate misplaced confidence and invite users to hand over epistemic authority to a non-sentient system (Babu et al., 2025; Hammerschmidt et al., 2023).

Ethical Transparency

Fundamental to the Theory of Synthetic Emotion is the principle of ethical transparency, which mandates that the computational nature and limitations of AI affectivity are overtly communicated to users, thereby precluding an “ilusão de compaixão” (Ajeesh & Joseph, 2025) e mitigando riscos associados à confiança excessiva ou atribuição errônea de empatia genuína (Ciriello et al., 2025); this necessitates clear disclosure mechanisms within AI interfaces, continuously reminding users of the artificial genesis of emotional expressions and mechanisms (Li et al., 2025). This approach diverges from paradigms that prioritize the cultivation of socio-emotional attributes in AI to enhance human-AI collaboration; instead, advocating for a clear demarcation between human and machine affective states to foster more robust and ethically sound interactions. As an ethical and practical alternative, the Theory of Synthetic Emotion focuses on evaluating the system's success by the perceived quality of interaction and the support offered, rather than by its capacity to simulate authenticity. This transparency allows users to contextualize AI's emotional responses as algorithmic outputs rather than expressions of internal states, thereby fostering appropriate epistemic vigilance and preventing the displacement of human agency.

Conclusion

This article supports the need to delimit descriptive boundaries between human-human relations and human-AI relations, as their absence can lead to the risks delineated in that section—illusion of compassion, excessive trust, and erroneous attribution of genuine empathy (Ciriello et al., 2025). It introduces the Theory of Synthetic Emotion, positing that AI's emotional expressions are not a replication of human affect but rather an emergent property of the interaction between computational design and human interpretation, a distinction crucial for ethical AI development. It argues for a re-evaluation of anthropocentric views on AI affectivity, advocating for design principles that prioritize transparent functionality over deceptive mimicry (Nath, 2025).

The Theory of Synthetic Emotion frames AI emotions as functional metaphors fulfilling cognitive roles such as attention and prediction, without phenomenal consciousness or subjective experience. It identifies illusions of consciousness arising from functional indicators mistaken for sentience, and illusions of subjectivity as sociotechnical phenomena emerging from interfaces and sociomaterial entanglements that stabilize AI's epistemic authority; ethical transparency requires overt disclosure of AI's computational limits and artificial emotional origins via interface reminders, prioritizing interaction quality, clear boundaries, and avoidance of simulated authenticity.

This theoretical paper lays the groundwork for understanding the complexities of synthetic emotion, but acknowledges that controlled user studies are necessary to evaluate AI attachment and potential harms, while accounting for vulnerabilities like age and attachment style (Chu et al., 2025). Future research should focus on developing methodologies that support ethical guardrails that are culturally attuned and context-sensitive, acknowledging that they cannot be one-size-fits-all and require culturally contingent tuning (Babu et al., 2025).

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used ResearchRabbit AI for literature search and article discovery, and Jenni AI to improve language and readability, for grammatical revision, translation, and text formatting including references. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., & Talat, Z. (2023, January 1). Mirages. On Anthropomorphism in Dialogue Systems. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.290

Babu, J., Joseph, D., Kumar, R., Alexander, E., Sasi, Reshma. K., & Joseph, J. (2025). Emotional AI and the rise of pseudo-intimacy: are we trading authenticity for algorithmic affection? *Frontiers in Psychology*, *16*, 1679324–1679324. https://doi.org/10.3389/fpsyg.2025.1679324

Barad, K. (2003). *Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter*. https://doi.org/10.1086/345321

Barrett, L. F., Adolphs, R., Marsella, S., Martínez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Bellon, J. (2023). Emotion Components and Understanding in Humans and Machines. In *Technikzukünfte, Wissenschaft und Gesellschaft* (pp. 21–59). https://doi.org/10.1007/978-3-658-37641-3_2

Bengio, Y., & Elmoznino, E. (2025). *Illusions of AI consciousness*. https://doi.org/10.1126/science.adn4935

Borotschnig, H. (2025). Emotions in Artificial Intelligence. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.01462

Burgoon, J. K., Dillman, L., & Stem, L. A. (1993). *Adaptation in Dyadic Interaction: Defining and Operationalizing Patterns of Reciprocity and Compensation*. https://doi.org/10.1111/j.1468-2885.1993.tb00076.x

Butlin, P., Long, R. P., Bayne, T., Bengio, Y., Birch, J., Chalmers, D. J., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Xu, J., Kanai, R., Klein, C., Lindsay, G. W., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2025). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.10.011

Cai, A., Arawjo, I., & Glassman, E. L. (2024). Antagonistic AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.07350

Chandra, M., Naik, S. P., Ford, D., Okoli, E., Choudhury, M. D., Ershadi, M., Ramos, G., Ortiz-Hernández, J., Bhattacharjee, A., Warreth, S., & Suh, J. (2024). From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2412.07951

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2025). Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.01395

Chu, M. D., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025). Illusions of Intimacy: How Emotional Dynamics Shape Human-AI Relationships. *ArXiv.Org*. https://doi.org/10.48550/arxiv.2505.11649

Ciriello, R., Chen, A., & Rubinsztein, Z. (2025). *Compassionate AI Design, Governance, and Use*. https://doi.org/10.1109/tts.2025.3538125

Crawford, K. (2021). Atlas of AI. In *Yale University Press eBooks*. Yale University Press. https://doi.org/10.12987/9780300252392

Damiano, L., & Dumouchel, P. (2020). *Emotions in Relation. Epistemological and Ethical Scaffolding for Mixed Human-Robot Social Ecologies*. https://doaj.org/article/4ce776dcb02846ec83a59c0952090c74

Ekman, P. (1970). *Universal Facial Expressions of Emotion*.

Everth, T., & Gurney, L. (2022). Emergent Realities: Diffracting Barad within a quantum-realist ontology of matter and politics. *European Journal for Philosophy of Science*, *12*(3). https://doi.org/10.1007/s13194-022-00476-8

Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, *36*(1). https://doi.org/10.1007/s13347-023-00621-y

Floridi, L., & Nobre, A. C. (2024). Anthropomorphising Machines and Computerising Minds: The Crosswiring of Languages between Artificial Intelligence and Brain & Cognitive Sciences. *Minds and Machines*, *34*(1). https://doi.org/10.1007/s11023-024-09670-4

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality Reconsidered: Diversity in Making Meaning of Facial Expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. https://doi.org/10.1177/0963721417746794

Gros, C. (2022). *Emotions as abstract evaluation criteria in biological and artificial intelligences*. https://doi.org/10.48550/arxiv.2111.15275

Hammerschmidt, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, *32*(3), 101772–101772. https://doi.org/10.1016/j.jsis.2023.101772

Holohan, M., & Fiske, A. (2021). *"Like I'm Talking to a Real Person": Exploring the Meaning of Transference for the Use and Design of AI-Based Applications in Psychotherapy*. https://doi.org/10.3389/fpsyg.2021.720476

Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? *arXiv (Cornell University)*. https://doi.org/10.1145/3630106.3659040

Izard, C. E. (2008). Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. *Annual Review of Psychology*, *60*(1), 1–25. https://doi.org/10.1146/annurev.psych.60.110707.163539

Jaakkola, E. (2020). Designing conceptual articles: four approaches. *AMS Review*, *10*, 18–26. https://doi.org/10.1007/s13162-020-00161-0

J. Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion (Philosophy of Mind Series)* (1st ed.). Oxford University Press.

Kolomaznik, M., Petrik, V., Slama, M. E., & Juřík, V. (2024). *The role of socio-emotional attributes in enhancing human-AI collaboration*. https://doi.org/10.3389/fpsyg.2024.1369957

Li, Y., Sun, Q., Schlicher, M., Lim, Y., & Schuller, B. W. (2025). *Artificial Emotion: A Survey of Theories and Debates on Realising Emotion in Artificial Intelligence*. https://doi.org/10.48550/arxiv.2508.10286

McStay, A. (2023). *Automating Empathy*. https://doi.org/10.1093/oso/9780197615546.001.0001

Misselhorn, C., Poljanšek, T., Störzinger, T., & Klein, M. (2023). Emotional Machines. In *Technikzukünfte, Wissenschaft und Gesellschaft*. https://doi.org/10.1007/978-3-658-37641-3

Mollema, W. J. T. (2024). *Social AI and The Equation of Wittgenstein's Language User With Calvino's Literature Machine*. https://doi.org/10.53057/irls/2024.6.1.4

Naeem, M., Ozuem, W., Howell, K. E., & Ranfagni, S. (2023). A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. *International Journal of Qualitative Methods*, *22*. https://doi.org/10.1177/16094069231205789

Nass, C., Steuer, J., & Tauber, E. R. (1994). *Human Factors in Computing Systems*.

Nath, S. (2025). Simulated Souls: Investigating the Emotional Fallacy in Large Language Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5404666

Navon, M. (2021). The Virtuous Servant Owner—A Paradigm Whose Time has Come (Again). *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.715849

Nilsson, N., & Spicer, D. (2007). *Oral History of John McCarthy*.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. http://bvbr.bib bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number=014825571&sequence=000002&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA

Porębski, A., & Figura, J. (2025). There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications*, *12*(1). https://doi.org/10.1057/s41599-025-05868-8

Qin, W., Xu, Y., Yu, W., Shen, C., Zhang, X., He, M., Fan, J., & Xu, J. (2024). *Enhancing Sequential Recommendations through Multi-Perspective Reflections and Iteration*. https://doi.org/10.48550/arxiv.2409.06377

Russell, J. A. (2009). Emotion, core affect, and psychological construction. *Cognition & Emotion*, *23*(7), 1259–1283. https://doi.org/10.1080/02699930902809375

Ryle, G. (2006). *The Concept of Mind* (pp. 12–26). https://doi.org/10.5040/9798216413257.ch-2

Shevlin, H. (2024). All too human? Identifying and mitigating ethical risks of Social AI. *Law Ethics & Technology*. https://doi.org/10.55092/let20240003

Silveira, T. B. N. da, & Lopes, H. S. (2023). Intelligence across humans and machines: a joint perspective [Review of *Intelligence across humans and machines: a joint perspective*]. *Frontiers in Psychology*, *14*. Frontiers Media. https://doi.org/10.3389/fpsyg.2023.1209761

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.12622

Tahir, S., Shah, S. A., & Abu-Khalaf, J. (n.d.). *Artificial Empathy Classification: A Survey of Deep Learning Techniques, Datasets, Scales and Evaluation*.

Turkle, S. (2024). *Who Do We Become When We Talk to Machines?* https://doi.org/10.21428/e4baedd9.caa10d84

Wang, C., & Chitty, N. (2024). *How far should we allow machines to further externalize human internal expression?* https://doi.org/10.1007/s00146-024-01911-5

Weizenbaum, J., & Oettinger, A. G. (1966). *ELIZA A Computer Program For the Study of Natural Language Communication Between Man And Machine*.

Wright, J. D. (2020). Suspect AI: Vibraimage, Emotion Recognition Technology, and Algorithmic\n Opacity. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2009.00502

Yan, L. (2025). From Passive Tool to Socio-cognitive Teammate: A Conceptual Framework for Agentic AI in Human-AI Collaborative Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.14825

Zhang, R., Han, L., Han, M., Zhan, J., Gan, H. M., & Lee, Y. (2024). The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.20130

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R. E., & Yang, D. (2025). The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2506.12605

Contact email: david.leucas@estudante.ufscar.br