# Otheroids or Anthropomorphism?

# An Empathy-Based Approach to Artificial Agents

Abootaleb Safdari[1]

[1] University of Bremen, Bibliothekstraße 1, 28359 Bremen, Germany
asafdari@uni-bremen.de

**Abstract.** The increasing presence of artificial agents (AAs) in everyday life has foregrounded a profound paradox in human-machine interaction: while people instinctively engage with AAs as if they possess consciousness and emotions, they often intellectually deny these very attributions. This contradiction has led to a dominant research paradigm, which this paper terms the "deception strategy," that dismisses such empathic behavior as a fallacy rooted in anthropomorphic illusions. This paper argues that this view is flawed, as it relies on (1) a rigid ontological divide between humans and machines and a (2) simplistic distinction between appearance and reality. Drawing from phenomenologically inspired enactivism, this paper proposes an alternative framework that reinterprets these empathic responses not as deceptive projections, but as the constitutive elements of a new form of social relation. By introducing the concept of the "otheroid," this paper offers a novel category for artificial entities that are experienced as neither fully human nor purely mechanical, thereby embracing the dynamic, reciprocal, and embodied nature of our interactions in an increasingly technologically mediated world.

**Keywords:** Anthropomorphism, Deception, Empathy, Phenomenology, Enactivism; Otheroid

## 1 Introduction

Our new brave world is increasingly populated by artificial agents (AAs), from robotic companions and digital assistants to autonomous systems in healthcare and industry. As these AAs become more sophisticated, human interaction with them has evolved beyond simple utility to encompass social and emotional dimensions. Studies have shown that people form strong, human-like bonds with these agents. For instance, research on the robotic seal PARO, used for dementia care, found that patients treated it as a social companion, a source of happiness, and even a conversation partner (Hung et al., 2021). Similarly, users of the robot Vector during quarantine described it as a family member and companion, engaging in shared activities like dancing (Odekerken-Schröder et al., 2020). This deeply ingrained human-like engagement with AAs has become a central topic of study across various disciplines, yet it suffers from significant conceptual ambiguity (Damholdt et al., 2023; Thellman et al., 2022).

This paper adopts the term anthropomorphism to refer to this phenomenon of attributing human-like qualities and mental states to AAs (Breazeal, 2002). The literature reveals a core paradox at the heart of this concept: while people readily anthropomorphize AAs during real-time, embodied encounters, they just as readily deny any true animacy or subjectivity to them upon detached, cognitive reflection. This apparent contradiction has led many researchers to adopt what this paper identifies as a deception strategy, arguing that such anthropomorphic behaviors are fundamentally illusory and should be discouraged. This paper will argue that this denial strategy is a conceptual dead end, rooted in a problematic adherence to a strict human-machine dualism and a simplistic distinction between appearance and reality. It will then propose a positive alternative: an empathy-based framework grounded in phenomenological enactivism and relationalism that reframes empathic human-AA interactions as genuine and constitutive of a new kind of social relation, which we describe with the new concept of the otheroid.

## 2 The Paradox and Deception Strategy

The paradox of human-AA interaction lies in the disconnect between implicit and explicit mind ascription. As Epley et al. (2007) noted, individuals with less time or cognitive resources for "effortful correction" are more likely to exhibit anthropomorphic behavior. This has been confirmed by empirical studies. In a study by Banks (2020), participants implicitly mentalized robots in a manner similar to humans during Theory of Mind (ToM) tests, yet explicitly denied that the robots had minds when directly asked. This suggests that the brain automatically processes social cues from AAs, but a reflective, conscious process overrides this intuitive response with the rational understanding that the robot is "just a machine." A similar finding by Fussell et al. (2008) showed that while participants used human-like language and attributed social traits to a robotic interviewer in spontaneous, real-time descriptions, they later explicitly denied that the robot possessed emotions or moods in a post-task survey. The data indicates that people engage in automatic social cognition with robots, but this is a far cry from a reflective belief that these agents have minds.

This paradoxical dynamic has led many scholars to conclude that attributing subjectivity to AAs is fundamentally mistaken—a form of deception. The deception strategy, as articulated by Matthias (2015), Placani (2024), Sharkey & Sharkey (2012, 2021), Sparrow & Sparrow (2006) and Winkle et al. (2021), posits that anthropomorphic design, emotional cues, and natural language in robots are deceptive techniques. They argue that these features exploit human social responsiveness without any corresponding internal states in the robot. This is seen as a form of "passive deception" arising from a performative gap between a robot's appearance and its actual function (2015). This view implies a duty to "see the world as it is" and avoid the "sentimentality" of believing an electronic toy is a friend (Sparrow & Sparrow, 2006, p. 155).

This widespread rejection of anthropomorphism as deception is, however, built upon two problematic assumptions. The first is a rigid ontological divide between humans (sentient, subjective beings) and AAs (passive, inanimate objects). The second is a

metaphysical separation between reality and appearance, where the robot's "true nature" as a machine is hidden beneath a deceptive, human-like facade. Critically, both assumptions fall apart under closer scrutiny of human-AA interactions.

## 3 A Critique of the Deception Strategy

As it has been mentioned, the "deception strategy" for understanding anthropomorphism is based on two faulty ideas. First, A strict divide between humans and machines that sees humans as having true subjectivity and machines as passive, lifeless objects. Thus attributing human traits to a machine is considered a mistake. Second, A separation of appearance and reality that suggests that a robot's human-like appearance is a fake illusion that hides its true mechanical nature. I believe both of these assumptions are problematic. Before addressing them in the next section, I would like to raise a preliminary concern about this strategy. First, even knowledgeable experts like Sherry Turkle (2011, p. 84) who know that a robot is "just code," still find themselves reacting to it as if it were a person. This proves that anthropomorphic behavior is not simply a mistake that can be corrected with knowledge. Second, The Roomba vacuum, a clearly mechanical and non-humanoid device, still inspires "Roombarization" in its users, who name it and express concern for it. This shows that anthropomorphism isn't just a response to a deceptive, human-like appearance; it's a deeper, more fundamental aspect of human-machine interaction.

### 3.1 Human-Machine Divide

In *Facing Gaia* (2015), Bruno Latour critiques entrenched dichotomies such as nature/culture, human/nonhuman, and animate/inanimate. He shows that natural entities are often treated as agents in practice—even by those who deny them agency. For example, the Mississippi River is described as “choosing” its path or “defying” dams. Such language reflects a practical recognition that the river exerts force, resists control, and requires negotiation, much like a political actor. Latour urges abandoning these binaries in favor of a more open notion of the “world” or “pluriverse” (after William James), which embraces the diversity and irreducibility of beings and resists fixed categories.

In this view, we live in a *metamorphic zone*—a space of constant transformation where boundaries between entities are fluid. Here, humans, animals, machines, and natural systems continually shape one another. It is within the metamorphic zone that anthropomorphism reveals itself not as a one-way projection from humans onto machines, but as a reciprocal dynamic, in which machines also shape and reconfigure human perception and embodiment. In *Seeing Like a Rover*, Janet Vertesi introduces the concept of technomorphism to describe this reversal of perspective, where humans begin to adapt to and internalize the robotic body’s way of engaging with the world (2015). Instead of simply imagining the robot as human-like, scientists and engineers working

on the Mars Rover missions learn to "see like a rover"—they embody its sensory constraints, operational logic, and limited range of motion.

Apart from these broader observations, we can narrow our focus to a specific subzone within the larger metamorphic zone—namely, the human–AA subzone—and develop it in a more systematic and conceptually rigorous way. To this aim, relationalism offers a more systematic way to understand human-AA interactions. This philosophical view argues that relationships between entities are fundamental, meaning an entity's nature is defined by its interactions, not by its fixed, intrinsic properties. This approach challenges traditional views that see subjects and objects as separate. In the context of robotics and AI, relationalism suggests that these systems are best understood and evaluated through their relationships with humans, society, and the environment, rather than as isolated tools (Coeckelbergh, 2012b; Darling, 2021; Gellers, 2021; Gunkel, 2023; Jones, 2013; Puzio, 2024).

### 3.2 Relationalism

Coeckelbergh (2010, 2012b, 2022) argues that traditional ontological, properties-based approaches treat moral status as something grounded in an entity's intrinsic characteristics. On this view, entities qualify for moral consideration if they possess certain defining traits: rationality, consciousness, sentience, or the ability to experience pleasure and pain. He then argues that this framework is fundamentally flawed in three ways:

1. **Epistemological problem** – There is no agreement on what *Q* should be; philosophers and ethicists disagree on which property is decisive.
2. **Detection problem** – Even if we agreed on *Q*, we often cannot reliably detect it, especially when it involves internal states.
3. **Continuum problem** – Most candidate properties come in degrees, making moral boundaries arbitrary and exclusionary.

In place of this intrinsic-property model, Coeckelbergh advances a *relational* approach. Here, moral status does not reside *in* the entity as an inherent essence. Instead, it emerges through lived interaction—how the entity is experienced, engaged with, and situated within broader social, cultural, linguistic, and spatial contexts. This "phenomenological–transcendental" shift reframes moral status as a product of relational dynamics rather than static ontological facts.

David Gunkel (2012, 2023), like Coeckelbergh, critiques the traditional approach to robot ethics, which frames the debate around dualist dichotomies like "person/thing." He argues this debate is a dead end because both sides rely on a flawed logical framework: that an entity must possess a certain quality (Q) to be considered a person. Similarly, Gunkel identifies three problems with this logic: there's no agreement on what "Q" is (the determination problem), no way to reliably detect it (the epistemological problem), and any judgment about its presence becomes a subjective decision (the decision problem). To escape this predicament, Gunkel proposes a radical shift: ethics precedes ontology. Instead of a robot's inherent nature dictating how we should treat it, he argues that our ethical decision to respond to it—drawing on Levinasian ethics—is what constitutes its status. In this view, how we treat a robot determines what it is, rather than the other way around.

### 3.3 Appearance and Reality

Mark Coeckelbergh challenges the "deception" narrative of anthropomorphism by arguing it's based on a flawed Platonic metaphysics that creates a strict divide between appearance and reality (2011, 2012a, 2018). First, the Platonic model is static and overlooks the temporal, processual nature of human–robot interaction. In reality, what the designer knows as the *program's time*—its internal sequences, pre-set triggers, and control logic—and what the user experiences as *interaction time*—the unfolding rhythm, pacing, and responsiveness—are not two separate realities. They are interwoven components of a single dynamic process in which technical execution and lived experience continually influence each other. Second, this process view shifts focus from opposing reality and illusion to recognizing two interrelated narratives—the designer's and the user's—which evolve together and shape the course of the interaction. Third, Coeckelbergh replaces the reality/illusion framework with that of performance, framing the encounter as co-performance in which users are active participants in constructing meaning, not passive recipients of preprogrammed output. Ultimately, he suggests using phenomenology to understand and go beyond this platonic metaphysics, and this leads to my positive proposal.

## 4 Phenomenological Alternative

From a phenomenological perspective, reality is not a static, external fact but emerges from the dynamic interplay between a subject and the world. This happens through the body, which acts as a bridge between our inner experience and our environment. A core idea in this perspective is intentionality, which means our conscious experience is always directed toward something outside of ourselves. This outward-directed quality is fundamentally tied to the body. The body provides us with motor intentionality, a non-conscious, skillful way of engaging with our surroundings. For example, when you instinctively know how to reach for a cup without deliberately thinking about it, that's motor intentionality at work. In this reciprocal exchange, the world is actively disclosed to the body as a field of practical opportunities. Objects are manifested as affordances—direct invitations to act that are immediately perceived based on the body's capabilities. For instance a handle is structured for grasping, and a horizontal stretch of ground is available for walking. The world is experienced not as a collection of indifferent things, but as a set of features that physically complement the body's repertoire of skills. This creates a continuous, reciprocal connection between the body and the world, which Merleau-Ponty (2005) called an intentional arc. Our body is directed toward the world, and the world responds by offering possibilities for embodied actions.

This embodied engagement gives rise to a relational conception of reality. This view distinguishes between an object's categorical properties (its fixed, viewpoint-independent features, like a plate's circular shape) and its perspectival properties (how it appears from a specific, embodied viewpoint, like a plate appearing elliptical when seen from

an angle) (Noë, 2004). A key point here is that the elliptical appearance isn't an illusion or a mistake. Perception, therefore, is not merely the passive registration of fixed, factual properties; it is the active grasp of how appearances vary as we move and relate to our surroundings. To perceive an object adequately is to be attuned not only to what it is, but to how it manifests across multiple perspectives. In this sense, perception is a skillful, embodied activity through which reality is enacted.

This approach offers a powerful new way to understand human-artificial agent relationships. Instead of asking whether our emotional experiences with a robot are "real" or "fake," this perspective suggests we should focus on the dynamic patterns of our embodied interactions. The reality of the relationship takes shape within these real-time, reciprocal engagements.

### 4.1 Human–AA Relations: The Structure of Interaction

In their study of the social robot Pepper, Ujike et al. (2019) observed varied interactions—conversation, eye contact, motivational cues, gestures, and more. A phenomenological approach to human-robot interaction suggests that to truly understand the dynamics at play, we need to shift our focus from the "what-ness" to the "how-ness" of the interaction. Instead of simply cataloging the content of the interactions—like what was said or what actions were performed—this perspective emphasizes the underlying processes and patterns. The most fundamental of these patterns is the action–response cycle, a continuous feedback loop where both the human and the AA are active participants. The cycle begins with one party initiating an action, and the other responding. This response then becomes a new action that the first party reacts to. In the case of the social robot Pepper, it would start with a greeting and a gesture. The patient's initial, sometimes non-verbal, responses—like gazing or a slight smile—are then met by Pepper's adjusted actions, such as verbal praise or new instructions. This ongoing cycle, where each participant's actions are influenced by the other's responses, creates a dynamic, relational interaction. This mutual adjustment produces a smooth, adaptive flow—what phenomenologists call *harmonious interaction*.

Harmonious interaction is crucial to the most fundamental level of experiencing another being as an other (a minded creature)—the *that*-level (Zahavi, 2014)—where we grasp an entity as minded before determining its specific states (*what*) or reasons (*why*). Safdari (2024; 2021) argues that when an action–response loop is harmonious, our anticipations are continually fulfilled, and the entity ceases to be a mere object, becoming an *other*. In the case of artificial agents, this shift produces what he calls an *otheroid*: a non-human entity that, through embodied and reciprocal engagement, is experientially constituted as an *other*. In Pepper’s case, this structure fosters an empathic relation in which the robot is no longer perceived as an inanimate tool but as an *otheroid*.

## 5 Conclusion

In this paper, I have proposed that in order to fully understand anthropomorphic behavior toward AAs, we must take these behaviors seriously i.e. we should not simply dismiss them as mere deceptions or cognitive illusions. To achieve this, I suggest that we turn to the concept of empathy. Accordingly, such behaviors are not misguided or erroneous projections of human traits onto machines, but rather constructive empathic relations. This shift in perspective can help us overcome a long-standing confusion within the literature on anthropomorphism.

The source of this confusion lies in a discrepancy: people tend to ascribe human-like traits to AAs during real-time, embodied interaction, yet resist or even reject such ascriptions when reflecting upon AAs in a detached, reflective manner. According to the framework I propose, this apparent contradiction is not a flaw or inconsistency in human cognition, but rather a constitutive and meaningful aspect of our evolving relationship with AAs.

From this perspective, the empathic relations we establish with AAs in pre-reflective, embodied encounters elevate them from mere objects to what I call otheroids—entities that occupy a relational space between objects and persons. Our interactions with otheroids are characterized by a dual structure: on the one hand, there is a spontaneous, pre-reflective empathic engagement; on the other hand, when we shift into a reflective and abstract mode, this empathic connection becomes unavailable or is actively suppressed. Far from being a problem to be solved, this duality opens up a new way of understanding our relationship with AAs—not as a deficient imitation of human-human relations, but as the emergence of a distinct and positive relational category. The concept of otheroids thus provides a framework through which we can explore and articulate the unique forms of sociality made possible by our encounters with artificial agents.

## References


1. Banks, J. (2020). Theory of Mind in Social Robots: Replication of Five Established Human Tests. *International Journal of Social Robotics*, *12*(2), 403–414. https://doi.org/10.1007/s12369-019-00588-x
2. Breazeal, C. L. (2002). *Designing Sociable Robots*. MIT Press.
3. Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5
4. Coeckelbergh, M. (2011). You, robot: on the linguistic construction of artificial others. *AI & SOCIETY*, *26*(1), 61–69. https://doi.org/10.1007/s00146-010-0289-z
5. Coeckelbergh, M. (2012a). Are Emotional Robots Deceptive? *IEEE Transactions on Affective Computing*, *3*(4), 388–393. https://doi.org/10.1109/T-AFFC.2011.29
6. Coeckelbergh, M. (2012b). *Growing Moral Relations Critique of Moral Status Ascription*. Palgrave Macmillan US.
7. Coeckelbergh, M. (2018). How to describe and evaluate “deception” phenomena:

recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology*, *20*(2), 71–85. https://doi.org/10.1007/s10676-017-9441-5

8. Coeckelbergh, M. (2022). Three Responses to Anthropomorphism in Social Robotics: Towards a Critical, Relational, and Hermeneutic Approach. *International Journal of Social Robotics*, *14*(10), 2049–2061. https://doi.org/10.1007/s12369-021-00770-0
9. Damholdt, M. F., Quick, O. S., Seibt, J., Vestergaard, C., & Hansen, M. (2023). A Scoping Review of HRI Research on 'Anthropomorphism': Contributions to the Method Debate in HRI. *International Journal of Social Robotics*, *15*(7), 1203–1226. https://doi.org/10.1007/s12369-023-01014-z
10. Darling, K. (2021). *The New Breed How to Think About Robots*. Penguin.
11. Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
12. Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. *Proceedings of the 3rd International Conference on Human Robot Interaction - HRI '08*, 145. https://doi.org/10.1145/1349822.1349842
13. Gellers, J. C. (2021). *Rights for Robots Artificial Intelligence, Animal and Environmental Law*. Routledge.
14. Gunkel, D. J. (2012). *The Machine Question Critical Perspectives on AI, Robots, and Ethics*. MIT Press.
15. Gunkel, D. J. (2023). *Person, Thing, Robot A Moral and Legal Ontology for the 21st Century and Beyond*. MIT Press.
16. Hung, L., Gregorio, M., Mann, J., Wallsworth, C., Horne, N., Berndt, A., Liu, C., Woldum, E., Au-Yeung, A., & Chaudhury, H. (2021). Exploring the perceptions of people with dementia about the social robot PARO in a hospital setting. *Dementia*, *20*(2), 485–504. https://doi.org/10.1177/1471301219894141
17. Jones, R. A. (2013). Relationalism through Social Robotics. *Journal for the Theory of Social Behaviour*, *43*(4), 405–424. https://doi.org/10.1111/jtsb.12016
18. Latour, B. (2015). *Facing Gaia: Eight Lectures on the New Climatic Regime* (C. Porter (trans.)). Polity Press.
19. Matthias, A. (2015). Robot Lies in Health Care: When Is Deception Morally Permissible? *Kennedy Institute of Ethics Journal*, *25*(2), 169–162. https://doi.org/10.1353/ken.2015.0007
20. Merleau-Ponty, M. (2005). *Phenomenology of Perception* (C. Smith (trans.); Taylor and). Routledge. https://doi.org/10.4324/9780203994610
21. Noë, A. (2004). *Action in Perception*. The MIT Press.
22. Odekerken-Schröder, G., Mele, C., Russo-Spena, T., Mahr, D., & Ruggiero, A. (2020). Mitigating loneliness with companion robots in the COVID-19 pandemic and beyond: an integrative framework and research agenda. *Journal of Service Management*, *31*(6), 1149–1162. https://doi.org/10.1108/JOSM-05-2020-0148
23. Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, *4*(3), 691–698. https://doi.org/10.1007/s43681-024-00419-4
24. Puzio, A. (2024). Not Relational Enough? Towards an Eco-Relational Approach in Robot Ethics. *Philosophy & Technology*, *37*(2), 45. https://doi.org/10.1007/s13347-

024-00730-2

25. Safdari, A. (2024). Toward an empathy-based trust in human-otheroid relations. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-02155-z
26. Safdari Sharabiani, A. (2021). Genuine empathy with inanimate objects. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-020-09715-w
27. Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, *14*(1), 27–40. https://doi.org/10.1007/s10676-010-9234-6
28. Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, *23*(3), 309–316. https://doi.org/10.1007/s10676-020-09573-9
29. Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, *16*(2), 141–161. https://doi.org/10.1007/s11023-006-9030-6
30. Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *ACM Transactions on Human-Robot Interaction*, *11*(4), 1–51. https://doi.org/10.1145/3526112
31. Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
32. Ujike, S., Yasuhara, Y., Osaka, K., Sato, M., Catangui, E., Edo, S., Takigawa, E., Mifune, Y., Tanioka, T., & Mifune, K. (2019). Encounter of Pepper-CPGE for the elderly and patients with schizophrenia: an innovative strategy to improve patient's recreation, rehabilitation, and communication. *The Journal of Medical Investigation*, *66*(1.2), 50–53. https://doi.org/10.2152/jmi.66.50
33. Vertesi, J. (2015). *Seeing Like a Rover: How Robots, Teams, and Images Craft Knowledge of Mars*. The University of Chicago Press.
34. Winkle, K., Caleb-Solly, P., Leonards, U., Turton, A., & Bremner, P. (2021). Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 101–109. https://doi.org/10.1145/3434073.3444666
35. Zahavi, D. (2014). Self and Other: Exploring Subjectivity, Empathy, and Shame. In *Oxford University Press*. https://doi.org/10.1017/CBO9781107415324.004