Who is a robot? A fundamental model of artificial identity

Lux Miranda
Department of Information Technology,
Uppsala University,
Uppsala, Sweden
lux.miranda@it.uu.se

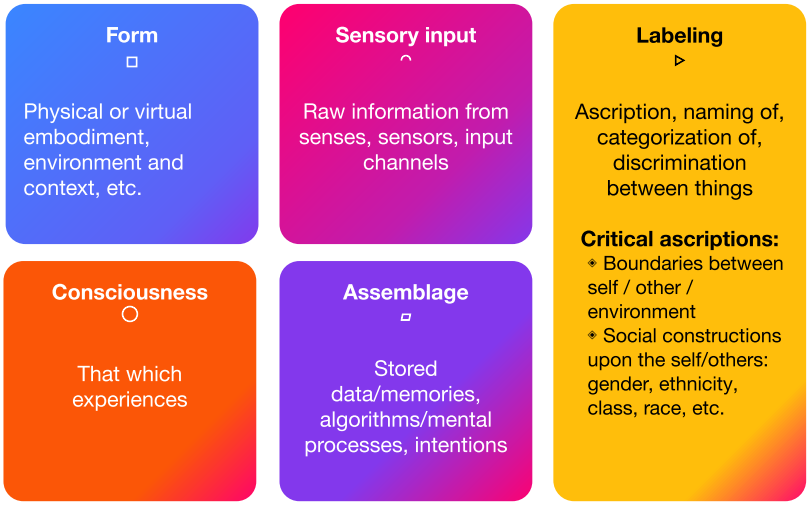

Figure 1: A summary of the five aggregates model.

ABSTRACT

Though some matters of consensus have begun to crystallize, scholars in human-robot interaction have thus far reasoned about artificial identity under many different definitions. Many of these seemingly disparate perspectives may, however, be unified into one coherent model through a synthesis of contemporary scientific and Buddhist philosophy of identity. Under this model, artificial and human identity are modeled equivalently under an assumption that there is no "immutable essence" which constitutes an agent's identity, but rather that identity may be defined as the sum of overlapping aggregates subject to change through time. The model reckons with the idea that much of what is conceived of as identity may be arbitrarily ascribed, artificial boundaries, but that these boundaries often constitute substantial social and psychological realities. This thinking is congruent with contemporary philosophical perspectives across disciplines from biology to cognitive science. The model may serve as a useful tool for roboticists to reason about identity in complex, dynamic situations, and provide a firm foundation for work which utilizes artificial identity. The model may even offer one or two possible answers to the question: Who is a robot?

Robo-Identity: Designing for Identity in the Shared World at HRI '24, March 11, 2024, Boulder, CO, USA

https://doi.org/10.5281/zenodo.10797808

KEYWORDS

artificial identity; robot identity; philosophy of HRI; five aggregates

ACM Reference Format:
Lux Miranda. 2024. Who is a robot? A fundamental model of artificial identity. In *Robo-Identity: Designing for Identity in the Shared World. Workshop at the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.5281/zenodo.10797808

1 INTRODUCTION

What is artificial "identity?" The word has been used by scholars of human-robot interaction (HRI) to mean at times very different concepts—personality, selfhood, group belonging, and other definitions still [14]—yet it remains intuitive that each of these things make up some facet of what we can call "identity." Through a synthesis of contemporary scientific and Buddhist philosophical ideas about the nature of identity, it is possible to unify these seemingly disparate perspectives into one coherent model. The model attempts to clear much of the fog surrounding what artificial identity is, how it works, and how it enmeshes with human identity; provide a firm philosophical foundation for knowledge which scholars of HRI have thus far intuited; and serve as a useful tool for reasoning about identity in complex, changing situations.

The basic premise of the model is as follows:

- Artificial identity can be modeled in fundamentally the same way as human identity. Distinction between the two are noteworthy exceptions rather than intrinsic differences.

- Boundaries defining different "identities" are ascribed arbitrarily but are often important psychological and social constructs.
- Different components of what is considered identity (such as a name, a body, or a consciousness) can be generally organized as belonging to five interacting and overlapping *aggregates* (Figure 1).
- The aggregates are temporally composable, accounting for the fact that identity may change through time.
- The aggregates are spatially composable, accounting for how boundaries separating identities are often freely redrawn to reason about identity at different levels of abstraction (Figure 2).

This model is congruous with perspectives across multiple disciplines. Similar reasoning has been applied in thinking on what constitutes an "organism" in light of cybernetic implications in biology [4], the nature of the "embodied mind" in cognitive science [22], hypotheses in complex systems concerning the emergence of life [23], phenomenological perspectives on identity in feminist theory [7, 8], and indeed ideas in HRI itself concerning thinking on identity in levels of abstraction [11], social identity theory [20], and malleable human-robot boundaries [15].

Modeling identity in this way frees us to confidently reason about it in a way that is useful and beneficial to us in our design, analysis, and advancement of artificial agents with identity. It resolves the question of "Who is a robot?" as, in one respect, unanswerable—given the false premise that there is a defined "who" with boundaries which can be deduced—and, in another respect, wholly modelable.

2 FIVE AGGREGATES

Awareness of the relevance of non-Western philosophy for the practice of AI and robotics has grown in recent years [9, 10, 13, 28]. Originating from Buddhist philosophy, the five aggregates (Sanskrit: *skandhas*) are, according to many interpretations, the five attributes which give rise to personhood and identity [6, 9, 10]. Reasoning with the aggregates, one posits that there is no "unchanging essence" which constitutes a person's identity, but that, like the ship of Theseus, it is instead defined as the sum of its ever-changing parts. The aggregates are metaphorically conceived of as "burning piles of phenomena"—a complex conglomeration of overlapping processes. Like piles of sand, they bleed into each other. Individual piles may be roughly discerned, but there are no hard boundaries between them; Bits of each aggregate may be found in every other aggregate.

Each aggregate may be intuitively identified as constituting some part of what is considered identity. For example, it is rather uncontroversial to consider the mind as one's identity. Others refute the Cartesian dualism imposing a separation between mind and body, and consider the mind-body as identity [7]. In other uses still, such as in literature concerning robots' "mind-body-identity mapping," the word "identity" refers to something separate from both the mind and the body [3, 12]. None of these interpretations are necessarily wrong, and they are all reconciled by the aggregates.

While each aggregate bears a traditional Sanskrit name, the choice of their English names is often the exercise of a given work's author as the Sanskrit names do not directly translate. The English names in this model depart from some of the usual selections—namely, we may specifically choose them to more clearly exhibit how each aggregate applies to both humans and machines. The five aggregates, with examples of each for both humans and machines, are:

1. **Form** (Sanskrit: *rūpa*)
   The physical or virtual manifestation of matter or other phenomena. Examples include the body, a virtual avatar, a voice, a chassis, facial features, etc. Consideration of form should always contextualize it in the environment, as identification may not always end in neat boundaries around a particular phenomenon. For example, one might extend identification with the organic body to include a prosthetic limb, or to the boundaries of one's car while at the wheel of it.

2. **Sensory input** (Sanskrit: *vedanā*)
   The raw information received from sensors, senses, input channels, etc. For humans, this encompasses the information from sensory organs such as the eyes, ears, skin, and so on. For machines, this typically takes the form of raw data from input channels or sensors such as cameras, microphones, tactile sensors, etc. Consider how someone with a particular chronic pain may come to identify with the pain itself, or how when we are ill we may not "feel like ourselves," identifying only with the usual feelings of non-illness as "me" and rejecting the feelings of illness as "not me." Naturally, it is always *Form* that is sensed.

3. **Labeling** (Sanskrit: *saṃjñā*)
   The naming and ascription of reified categories to sensory input; that is, the imposition of discrete labels upon continuous phenomena. For example, a human might draw artificial lines separating "red" from "orange" from "yellow" along the 600-700 nanometer portion of the visual light spectrum, or label continuous changes in psychosomatic sensations as discrete feelings of "joy" or "contentment." Machines do this very directly with machine learning, such as learning to distinguish objects in a video feed and labeling them as "car," "person," "bus," etc. As we will see in the next section, it is this process which leads to a great deal of what is considered identity—particularly through labeling the boundaries between "self" and "other" and the ascription of social categories such as race and gender.

4. **Assemblage** (Sanskrit: *saṅkhāra*)
   Mental or computational processes such as thoughts, programs, algorithms, memories, stored data, intentions, plans, etc. Identification can be seen as "I *am* my thoughts" or pointing to the same program which is running across multiple embodiments as the "identity" of an agent. See also the identification of self with memories in the form of, for example, "What is a man but the sum of his memories?" [2]

5. **Consciousness** (Sanskrit: *vijñāna*)
   That which cognizes and experiences. One can be aware of one's own thoughts, feelings, sensory input, and labelings,

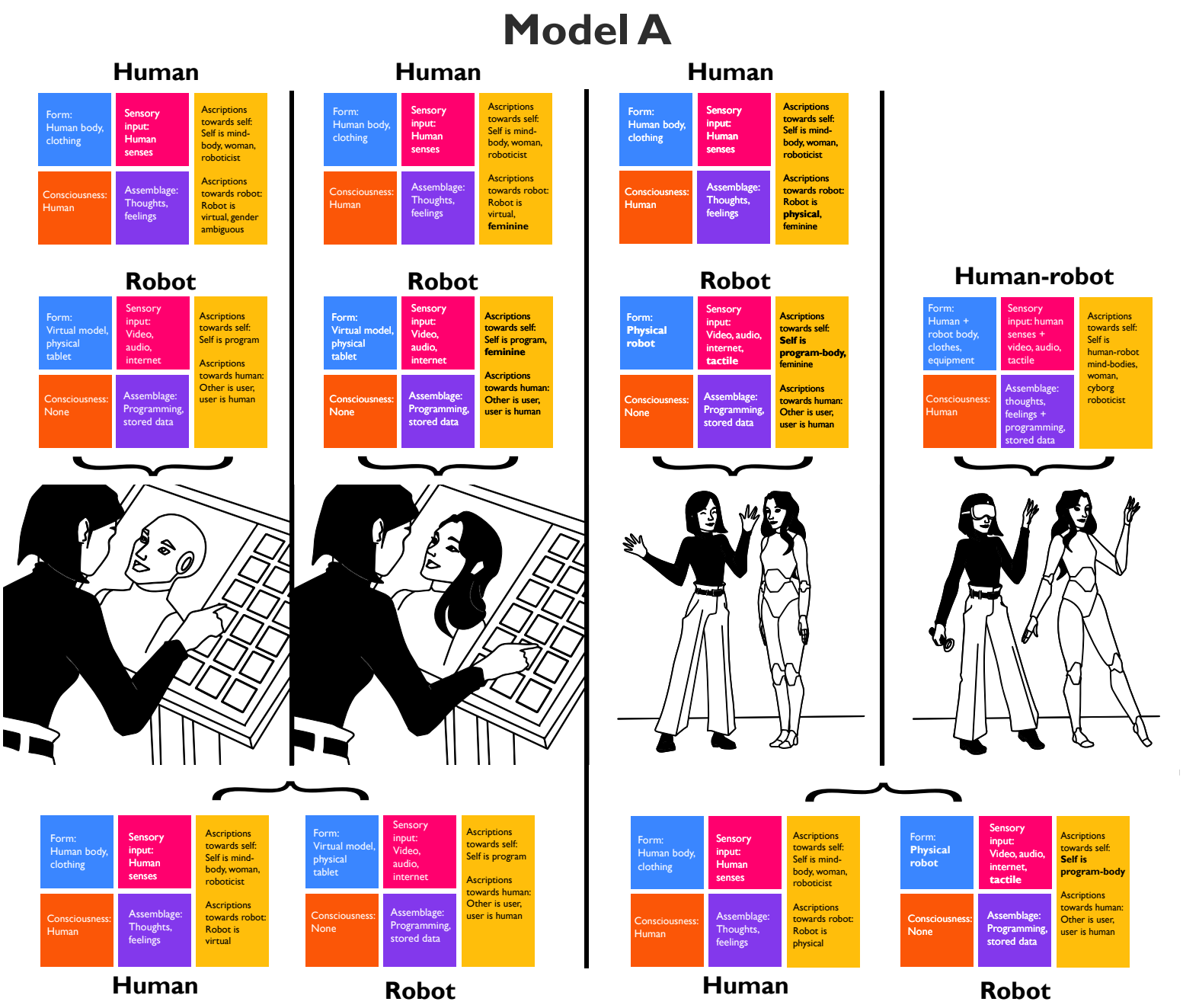

Figure 2: An example showcasing the temporal and spatial composability of the model: A user and robot's identity characteristics change with time. (1) The user begins interacting with an agent in virtual form on a tablet. (2) The user changes the agent's gender characteristics according to their preference. (3) The agent's identity is re-embodied into a physical robot. (4) The user takes manual control of the robot's motor functions using a VR headset, allowing them to also re-embody into the robot platform and blur the boundary between self and other. Two potential identity modelling solutions are presented. Model A uses smaller timesteps and treats the human and robot as a single agent in the final timestep. Model B uses larger timesteps and keeps the identity of the human and robot separated. Both models are equally valid choices of the modeller, and either may be used depending on the needs and goals of the modeller in identity design and identity analysis.
Illustration © 2024 Lauren Galvan

yet still not identify with them. Consciousness is what remains, and here is the *awareness* of these things. Naturally, human consciousness goes here. The model allows for the possibility of artificial consciousness, but it does not require it. Given that not all five aggregates must be identified with for an identity to exist, artificial identity still exists in the absence of artificial consciousness. An important note: Since the "amount of consciousness" plays a large role (but not the sole role) in how we determine the amount of ethical treatment required for a given being, consciousness is important to model for when considering, for example, human-robot hybrid identities or organic-synthetic cyborg organisms.

3 ASCRIPTION

Much of what is conceived of as identity are ascriptions under the model. Ascriptions, in this context, are artificial boundaries, names, and categories created through *Labeling*. These typically manifest as social or psychological constructs and encompass a wide variety of phenomena such as gender, the boundary between the body and environment, or the point in time at which an identity comes into existence. These are often attempts to model phenomena which exhibit genuine variation, but cannot be truly quantized in reality. Ascriptions can be understood as the answer to the question: What is reified?

Though technically artificial, ascriptions are not without systemic forces driving their creation and real-world consequences of their existence. Ascriptions which are agreed upon by multiple people, or are otherwise reified psychologically, fundamentally shape the landscape of our social and internal experience. The key is that these boundaries may be drawn and redrawn at any time and in any configuration.

Understanding ascriptions in this way allows us to work with them to our advantage. It allows us to know, for example, that all possible embodiment configurations of a robot's mind, body, and personality are equally valid and readily explained with the model, and that no effort is needed to locate the "who" in an engineered identity. In such an instance, it is users' own acceptance and understanding of the configuration that becomes the foremost concern. We may also, for example, phrase a question on whether robots may have "personal identity" [1] as equivalent to the question "may robots self-determine internal ascriptions of identity?"

It would be impossible to model every possible ascription, as a general model of them would essentially be a model of the whole of human experience. There are certain types of ascriptions, however, that we must turn our focus to for practical and ethical reasons. Namely: social identities which may relate to inequity, ascribed boundaries which create ideas such as a "self" which is separate from an "other," and ascriptions which occur across time.

3.1 Social identity

As opposed to *personal identity*, which is self-ascribed, *social identity* is socially constructed and socially ascribed [16, 20]. Given that social inequities arise along lines of social identity like race, gender, and class [5], it is paramount to always model them when applicable. These sorts of identities are regularly ascribed to agents such as digital assistants and robots—especially when human-like or viewed as social actors [17, 18, 20]. Works in HRI have made an increasingly extensive case on the importance of sensitivity towards these identities in interaction design for hindering the propagation of harmful norms and ceasing to uphold social inequity [16, 17, 21, 24, 26, 27]. Of course, less sensitive social identities (such as profession, fanhood for a particular rugby team, etc.) may also be modeled when useful. Designers must be careful to note that the intended social identities for an agent may not always align with the identities which users actually ascribe to the agent.

3.2 Spatial composability

Under the model, spatial boundaries between identities are ascribed. We may take a postphenomenological [19] approach wherein, rather than attempting to understand the identity of discrete individuals or agents, we understand identity as something with malleable boundaries that is mediated by the complex interactions between humans, technology, and the environment. For the purposes of our own analysis and modeling, we may draw or not draw these boundaries wherever it is useful to do so.

Examples of important ascribed spatial boundaries are the boundaries between self and other; between others; between self and environment; or between mind, body, and personhood. Treating the spatial aspect of identity in this way allows us to consider multiple levels of abstraction and reason about identity beyond the level of a single individual (such as the collective identity of a group).

If (and only if) a boundary between "self" and "other" is established, then we may speak of internal (self-directed) ascriptions versus external (other-directed) ascriptions. It is only under this condition that the questions arise: "Who am I?" and "Who are they?" When modeling, it is important to consider the *source* (or perspective) of these ascriptions.

3.3 Temporal composability

Temporal boundaries are also ascribed. That is, identity must always be modeled *within a particular window of time*. All aspects of identity—personality, embodiment, thoughts, social associations, etc.—are subject to constant change for both humans and machines. While this is generally a slow process for humans, artificial agents are exceptionally flexible in this regard [25] and can often alter major aspects of identity instantaneously (for example, facial, gender, and vocal cues of a Furhat robot). Traditionally, this is also an important property of not only ascriptions but the aggregates themselves [9]. The *composability* of this property arises when we can model key identity characteristics as they overlap with each other in time, depending on which characteristics we are interested in (Figure 2).

4 CONCLUSION

Thus the basic model is outlined. Further refinements on this model may be in order as it is put into practice and tested. Combined with the practical minimal modeling example in Figure 2, this information shall hopefully be sufficient as a useful tool for roboticists to disentangle even the most complex of artificial identity situations and allow us to confidently deploy artificial identity in our work.

ACKNOWLEDGEMENTS

Many thanks to Lauren Galvan for the illustration, to Tan Zhi Xuan for the manuscript feedback, and to the reviewers for their suggestions. Thanks also to the Wallenberg AI, Autonomous Systems and Software Program — Humanity and Society (WASP-HS) for valuable conversations and collegial support.

REFERENCES

[1] Marcos Alonso. 2023. Can Robots have Personal Identity? *International Journal of Social Robotics* 15, 2 (Feb. 2023), 211–220. https://doi.org/10.1007/s12369-022-00958-y

[2] Alexandre Begnoche. 2011. Assassin's Creed Revelations.

[3] Alexandra Bejarano, Sebastian Negrete-Alamillo, and Tom Williams. 2023. Conversations with Identity Performing Robots: Considerations for Algorithms and Interfaces. (2023).

[4] Wesley P Clawson and Michael Levin. 2022. Endless forms most beautiful 2.0: teleonomy and the bioengineering of chimaeric and synthetic organisms. *Biological Journal of the Linnean Society* (July 2022), blac073. https://doi.org/10.1093/biolinnean/blac073

[5] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality.* John Wiley & Sons. Google-Books-ID: fyrfDwAAQBAJ.

[6] Jay L. Garfield. 2014. *Engaging Buddhism: Why It Matters to Philosophy.* Oxford University Press. Google-Books-ID: UDnKBQAAQBAJ.

[7] Elizabeth Grosz. 2020. *Volatile Bodies.* Routledge, London. https://doi.org/10.4324/9781003118381

[8] Donna Haraway. 2006. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In *The Transgender Studies Reader.* Routledge. Num Pages: 16.

[9] Soraj Hongladarom. 2020. *The Ethics of AI and Robotics: A Buddhist Viewpoint.* Rowman & Littlefield. Google-Books-ID: MprzDwAAQBAJ.

[10] James Hughes. 2012. Compassionate AI and selfless robots: A buddhist approach. *Robot ethics: the ethical and social implications of robotics* (2012), 69–83. Publisher: Cambridge, Massachusetts: MIT Press.

[11] Ryan Jackson, Alexandra Bejarano, Katie Winkle, and Tom Williams. 2021. Design, Performance, and Perception of Robot Identity.

[12] Karla Bransky Kelly, Penny Sweetser Kyburz, Sabrina Caldwell, and Kingsley Fletcher. 2024. Mind-Body-Identity: A Scoping Review of Multi-Embodiment. ACM/IEEE. https://doi.org/10.1145/3610977.3634922 Accepted: 2024-01-09T23:36:16Z Last Modified: 2022-11-17.

[13] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion).* Association for Computing Machinery, New York, NY, USA, 10–18. https://doi.org/10.1145/3434074.3446908

[14] Minha Lee, Dimosthenis Kontogiorgos, Ilaria Torre, Michal Luria, Ravi Tejwani, Matthew J. Dennis, and Andre Pereira. 2021. Robo-Identity: Exploring Artificial Identity and Multi-Embodiment. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21 Companion).* Association for Computing Machinery, New York, NY, USA, 718–720. https://doi.org/10.1145/3434074.3444878

[15] Dominika Lisy. 2024. A Relatable Subject: Body Boundaries and Sensorial Hierarchies. In *Proceedings of the Fourth WASP-HS Winter Conference*, Ericka Johnson and Eva Sjöstrand (Eds.). Umeå, Sweden.

[16] Lux Miranda, Ginevra Castellano, and Katie Winkle. 2023. Examining the State of Robot Identity. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Stockholm Sweden, 658–662. https://doi.org/10.1145/3568294.3580168

[17] Lux Miranda, Ginevra Castellano, and Katie Winkle. 2024. A Case for Diverse Social Robot Identity Performance in Education. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Boulder, CO, USA, 8. https://doi.org/10.1145/3610978.3640768

[18] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology* 27, 10 (May 1997), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

[19] Robert Rosenberger and Peter-Paul Verbeek. 2015. A field guide to postphenomenology. *Postphenomenological investigations: Essays on human-technology relations* (2015), 9–41. Publisher: Lexington Books Lanham, MD.

[20] Katie Seaborn. 2022. From Identified to Self-Identifying: Social Identity Theory for Socially Embodied Agents. In *Proceedings of the IEEE/ACM HRI 2022 Workshop on Robo-Identity 2.*

[21] Michael Stolp-Smith and Tom Williams. 2024. More Than Binary: Transgender and Nonbinary Perspectives on Human Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.* Boulder CO USA.

[22] Francisco J. Varela, Evan Thompson, and Eleanor Rosch. 2017. *The Embodied Mind, revised edition: Cognitive Science and Human Experience.* MIT Press. Google-Books-ID: bxLiDQAAQBAJ.

[23] Sara Imari Walker and Paul CW Davies. 2013. The algorithmic origins of life. *Journal of the Royal Society Interface* 10, 79 (2013), 20120869. Publisher: The Royal Society.

[24] Tom Williams. 2023. The Eye of the Robot Beholder: Ethical Risks of Representation, Recognition, and Reasoning over Identity Characteristics in Human-Robot Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Stockholm Sweden, 1–10. https://doi.org/10.1145/3568294.3580031

[25] Katie Winkle, Ryan Jackson, Alexandra Bejarano, and Tom Williams. 2021. On the Flexibility of Robot Social Identity Performance: Benefits, Ethical Risks and Open Research Questions for HRI.

[26] Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. 2023. Feminist Human-Robot Interaction: Disentangling Power, Principles and Practice for Better, More Ethical HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Stockholm Sweden, 72–82. https://doi.org/10.1145/3568162.3576973

[27] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* ACM, Boulder CO USA, 29–37. https://doi.org/10.1145/3434074.3446910

[28] Tan Zhi-Xuan. 2020. AI Alignment, Philosophical Pluralism, and the Relevance of Non-Western Philosophy. https://www.alignmentforum.org/posts/jS2iiDPqMvZ2tnik2/ai-alignment-philosophical-pluralism-and-the-relevance-of