Defining the Essence of the Self in Exploring the Notion of Altruism and Establishing Trust in Human | Robot Interaction (HRI)

Dr. Hortense Gerardo, University of California, San Diego

Dr. Hortense Gerardo is a playwright, screenwriter, and anthropologist and serves as the Director of the Anthropology, Performance, and Technology (APT) Program at the University of California, San Diego. Her works have been performed nationally and internationally. She is a Co-founder of the Asian American Playwright Collective (AAPC), and head of the Screenwriting competition on the Board of the Woods Hole Film Festival. For more information go to: www.hortensegerardo.com

Dr. Brainerd Prince, Plaksha University

Brainerd Prince is Associate Professor and the Director of the Center for Thinking, Language and Communication at Plaksha University. He teaches courses such as Reimagining Technology and Society, Ethics of Technological Innovation, and Art of Thinking for undergraduate engineering students and Research Design for PhD scholars. He completed his PhD on Sri Aurobindo's Integral Philosophy from OCMS, Oxford – Middlesex University, London. He was formerly a Research Tutor at OCMS, Oxford, and formerly a Research Fellow at the Oxford Centre for Hindu Studies, a Recognized Independent Centre of Oxford University. He is also the Founding Director of Samvada International Research Institute which offers consultancy services to institutions of research and higher education around the world on designing research tracks, research teaching and research projects. His first book The Integral Philosophy of Aurobindo: Hermeneutics and the Study of Religion was published by Routledge, Oxon in 2017. For more information, please visit: https://plaksha.edu.in/faculty-details/dr-brainerd-prince

Mr. B. Lallian Ngura, Centre for Thinking Language and Communication (CTLC), Plaksha University

B. Lallianngura has completed post-graduate studies in philosophy from the University of Delhi. He is pursuing doctoral research in philosophy at IIT Bombay. He is a part of the research team at Centre for Thinking, Language and Communication at Plaksha University. His research focuses on the question of self and subjectivity and its relation to power-knowledge discourse in Michel Foucault.

Defining the Essence of the Self in Exploring the Notion of Altruism and Establishing Trust in Human | Robot Interaction (HRI)

Authors:

[Theory Paper, Ethics of Emerging Technology]

Abstract

Artificial Intelligence (AI) and cognitive robotics (CR) technologies are redefining and disrupting the way people work and live in many different domains. With an aging Baby Boomer generation, an increase in the small, nuclear family unit (as opposed to the multi-generational kinship assemblages housed under one roof), and a decrease in birth rate in so-called "developed" countries, there is an increasing trend in the use of these technologies to conduct personal care for aging populations and for the very young.[1] "Gerontechnology" based on Artificial Intelligence (AI) is expected to enable a predictive, personalized, preventive, and participatory elderly care". [2][3] As medical dependency on AI accelerates, we are confronted with issues of safety and trust around its use. This paper uses a literature review as a methodology by which to discern similarities and differences in definitions of the "Self" as applied to humans and in parlance around AI and CR. By refining the definition of what is meant from a philosophical perspective by the concept of the "Self," "Consciousness" and "Altruism" and juxtaposing these against the functional distinctions between Theory of Mind and Self-Aware AI, we posit the theoretical possibility, based on existing literature, of decision-making, self-aware AI capable of what might be considered a form of collective identity-based, altruistic behavior. This analysis is intended to inform considerations of the ethical implications to engineering of such systems in caring for the elderly and the young.

Introduction

"The use of robots in healthcare treatments (can be divided) into three categories: support, mitigation, and reaction" [4]. AI is a human intelligence simulation processed by machines to perform tasks, whereas Cognitive Robotics parallels cognitive science in its explorations of cognitive phenomena like learning, memory, reasoning, perception, and attention. In simple terms, a robot is a machine, and AI is the algorithmic complex that ignites perceptual abilities in a machine. Currently, there are four different types of AI:

1. Reactive Machines AI programmed with predictable output based on programmed input.
2. Limited Memory AI can acquire, adjust, and interpret data, and use prior experience to aid in decision-making processes.
3. Theory of Mind AI, the current state of advanced AI, are programmed with decision-making abilities that mimic humans.
4. Self-aware AI are perceived to be the most advanced form of AI that are self-aware of their internal state, emotions, behaviors, and reasoning. [5][6]

The engineering of these systems will require the development of ethical frameworks capable of supporting these evolving technologies. A common theme in pondering the consequences of developing Self-Aware AI is the question: will the AI decide humans are unnecessary or

destructive to the planet and destroy them? Or might they decide to aid and enhance the life of humans? [7] Earlier studies exploring issues of trust have been measured indirectly through game play and the element of surprise elicited from Reactive Machines AI and Limited Memory AI [8]. Assessing levels of trust might also be explored by studying to what extent AI exhibits "altruistic" behavior. Are cognitive, pre-trained AI truly capable of learning "altruistic" behaviors? To that end, we consider the definition of "altruism" in the context of AI and its implications with respect to human | robot interaction (HRI) and the development of trust. The latest generations of AI models show no signs of self-awareness according to fourteen parameters of human cognition. However, there is currently no theoretical barrier preventing AI from reaching self-awareness. [9]

This paper proposes first to examine the literature to review what has traditionally defined the concept of the human "Self" in philosophical terms, and second, to review the literature on the concept of "Self" in terms of Limited Memory and Theory of Mind AI. A brief compendium of these bodies of literature will allow for an informed query on the nature of "consciousness" in relation to AI, and how this is similar to or different from human consciousness.

Conceptual Framework - Literature Review as Methodology

In order to address the question of whether or not AI can harbor a sense of "Self" we examine the literature for various definitions of the "Self" as it is defined in philosophical terms for humans, as well as the literature that addresses the various types of AI and how these distinctions might lend itself to an elastic definition of the "Self."

Philosophical Notions of the Human "Self"

Charles Taylor on Self: Consciousness and the first-person perspective

Charles Taylor, the pioneer contemporary philosopher on the self or selfhood, claims in his classic work *Sources of the Self* that being a human agent or self entails the notions of inwardness, freedom, individuality, and being embedded in nature. [10, p. ix]

Taylor especially highlights the first-person perspective as one of the main markers of modern selfhood. This view involves taking up a reflexive stance, which necessitates an inward turn i.e., a turn towards one's self. [10 p.130] According to him, "there is a crucial difference between the way I experience my activity, thought, and feeling, and the way that you or anyone else does. This is what makes me a being that can speak of itself in the first person." [10 p. 131] Taylor calls the adoption of this inward first-person standpoint in knowledge as "a stance of radical reflexivity" [10 p.130] According to this view, every knowledge or awareness is always already that of an agent (or self). "The world as I know it is there for me, is experienced by me, or thought about by me, or has meaning for me." [10 p.130] This first-person standpoint sees the world from the perspective of the knowing or experiencing agent i.e. it focuses on what it is like to be an experiencing agent rather than on the things experienced. This viewpoint is radically different from the 'objective' standpoint i.e. the viewpoint of the natural sciences, which offers a "view from nowhere" and focuses only on the things or objects experienced. [10 p.130]

Michel Foucault on Self:

Disciplinary Self

Michel Foucault, one of the most influential modern philosophers on modern selfhood (or 'subjectivity' in his language), opines that the modern self is the product of knowledge and power relations. His exploration of the "Disciplinary Self" revolves around the concept that modern societies transform human beings into subjects through various forms of objectification, with a focus on the body as a target of power. [11] In his iconic work, *Discipline and Punish* (1977) he describes a system in which bodies are broken down, analyzed, and scrutinized to become subservient and efficient, effectively producing disciplined subjects that serve as both objects and instruments of power. [12] Foucault dissected the role of disciplinary systems, such as prisons and surveillance - as exemplified by the Panopticon - in molding individuals and their souls. These mechanisms enforce a dystopian state of constant visibility, leading to self-inhibition and the rise of power structures, suggesting that the soul is not immutable, but constructed through the process of control and punishment. [13] This analysis offers a critical insight into how authority operates in society, influencing identity and agency. [14]

The Human " Self" in Relation to Technologies of Power:

Foucalt describes his work as an historical analysis of how the systematics of knowledge relates to power and self regulation. He examined the state of insanity, not from a clinical point of view, but as a lens through which to study how societies manage individuals within asylums, as well as within the context of being in a state of being manipulated by external mechanisms as well as by self-imposed inhibitory practices. [14]

Artificial General Intelligence (AGI), Intelligent Agents (IA), Multi-Agent Systems (MAS) - Al Definitions of the "Self"

The field of Artificial Intelligence (AI) employs co-related terms such as Artificial General Intelligence (AGI) or "strong" AI, Intelligent Agents (IA), and Multi-Agent Systems (MAS) in order to engage with questions concerning AI Self and its relation to ethical principles, ethical reasoning, and responsibility. The emerging field of Intelligent Agents (IA) addresses issues of agency, autonomy, self-interest and so on. Likewise, the related field of Multi-Agent Systems (MAS) extends this framework to model interactions between autonomous agents and their emergent properties. MAS address questions of ethics and responsibility within the context of conflict of self-interests between disparate agents. [15, p. 3]

In theory, AGI is a type of program or model that has the full intellectual capabilities of a human, i.e. general intelligence. AGI would have abilities like reasoning, common sense, abstract knowledge, and creativity. Essentially it would be able to autonomously perform tasks without human instruction. Although true AGI does not exist yet, some experts believe it could be achieved in the near future. Strong AI is another term for AGI or artificial general intelligence. At present it is a purely theoretical form of artificial intelligence that would autonomously "think" and act like a human. However, it should be stressed that there is no consensus that AGI can be achieved in the near future. Some doubt that it may ever be achieved and contend that ChatGPT and other generative technologies have no knowledge of reality and are simply correlation technologies. [16] [20]

The AI Elastic Sense of “Self”

Srinivasa et al. propose that understanding and modeling the "elastic sense of self" is crucial for developing responsible artificial intelligence (AI). This concept refers to the ability of beings to extend their sense of self to external entities or concepts, effectively investing part of their biological and cognitive resources to support these identified objects. This mechanism, which can be observed in the way humans identify with their children, country, or causes, is linked to empathy and possibly to the mirror neuron system. In AI research, attempts to model a sense of self have focused on creating computational models that feature autonomy, intentionality tempered by beliefs and knowledge, and adaptability through reinforcement learning. The "elastic sense of self" is particularly highlighted as a potential foundation for innate responsibility and ethics in humans, suggesting that our identity can include, and extend to, a wider set of objects and concepts beyond our physical selves, fostering a sense of belonging and loyalty. [15]

Human Phenomenal vs Access Consciousness - Frameworks for Comparison to AI

Several authors have offered differing views on the development of conscious AI, ranging from advocating against it [16], to calling for a moratorium [17], supporting its development [18], and arguing for careful consideration of AI consciousness [19]. Consciousness, in this context, refers to the capacity for subjective experience or what it is like to have an experience. The report outlines a method for assessing AI consciousness based on computational functionalism, scientific theories of consciousness, and a theory-heavy approach. Indicators drawn from various theories of consciousness provide a rubric for evaluating the likelihood of consciousness in AI systems. The findings suggest that while current AI may not be conscious, the potential for developing conscious AI exists, given that most conditions for consciousness can be met with existing AI techniques.

Various theoretical approaches to measuring AI consciousness include but are not limited to the following: recurrent processing theory (RPT-1, RPT-2), Global workspace theory (GWT-1, GWT-2, GWT-3, GWT-4), Computational higher-order theories (HOT-1, HOT-2, HOT-3, HOT-4), Attention schema theory (AST-1), Predictive processing (PP), Agency and embodiment (AE-1, AE-2) [17, pp. 4-6].

Refining the Definition of Human Altruism to Ask: is AI capable of altruistic behavior?

The term, “altruism”, originating from French philosopher August Comte, emphasizes unselfish concern for the wellbeing of others, in contrast to egoism. The intention to benefit others without thought of personal gain are defining characteristics of this disposition. [18] The concept of

altruism varies across disciplines, from reproductive outcomes in biology that may benefit an agent, to philosophical or psychological factors that might motivate behaviors to help others.

[19] [23] Altruism may inform moral thinking as a duty to consider others' circumstances, but may also be the basis for a supererogatory insistence that may only benefit the few at the expense of the greater majority. [22]

"There are two components to altruism: positively, a concern for the interests of another person; negatively, a lack of concern with one's own interests." [18] Association with a collective identity, one which is not directly traceable to oneself, describes altruism of a negative characteristic. [18, p. 61]

Discussion

This brief but by no means comprehensive literature review of the "Self" in philosophical terms as it applies to humans reveals several contexts by which the term might pertain to AI and CR. Likewise, the various types of AI reveal the necessity of expanding the definition of what might constitute a "Self" in engineering terms.

A distinction between phenomenal versus access consciousness in humans provides a specific window by which the possibility of "consciousness" might pertain in terms of phenomenological input that AI can process and transform into manifestable output. Access consciousness by contrast appears to be of a higher order of AI and provides the boundary between Theory of Mind AI and Self-Aware AI. However, the literature also posits that the capability of creating conscious AI without radically new hardware is theoretically possible at present. [21] An application of one of the various definitions of the term "altruism," specifically, as it pertains in the case of pure collective identification might qualify certain behaviors generated by such AI and CR as "altruistic" in nature. Finally, a parsing between phenomenal versus access consciousness in humans paves the way for a type of consciousness - phenomenological in nature - that applies to Theory of Mind AI. These constitute the current state of advanced AI which are programmed with decision-making abilities that mimic humans. The connection between phenomenologically conscious or Theory of Mind AI and access conscious or Self-aware AI is as of this writing, open-ended but theoretically possible if not probable with current technology. These considerations are helpful in programming and engineering design of AI with respect to care for the elderly and the young. With the imminent transition of the Baby Boom Generation to elder care and the decrease in birth rates in developing countries, [24] the reliance on AI to care for these demographics will increase, necessitating a transition in society's relationship to AI from one of reliability to that of trust. Partner and social robots are already being widely used in contexts ranging from education, entertainment, therapy, and assistance. [25] The development of generative AI will further accelerate the integration of partner/social robots into daily life. In order for this integration to happen smoothly, it will be crucial to develop technologies that can foster positive human - robot interaction that lends itself to the development of trust.

Conclusion

This paper reviews the various ways that the concept of "Self" has been defined for humans, as well as the ways that different types of AI might yield different concepts of how a "Self" might

pertain. As the various definitions of "self" and "consciousness are refined, it is clear that we are approaching an asymptotic parallel between human and AI applications of these terms. The thought experiment of posing the possibility of AI being capable of altruistic behavior is approached here by distinguishing between various forms of altruism. By this method it is possible to apply this definition to AI to a select subset in the category termed Theory of Mind AI. The distinction between phenomenal consciousness and access consciousness would appear to be the difference between AI capable of decision-making, and one that is self-aware. Current technology already exists for AI to automatically shut off in close proximity to humans, or to signal malfunction that might pose a safety threat, for example, as part of common design practice in the field of engineering safety [26, p 798]. Further studies might include specific case studies of Theory of Mind AI demonstrating examples of collective identity altruistic behavior. While speculative future popular culture writings tend to lean heavily toward a dystopian future of nihilistic AI, it is also possible that next level Self Aware AI may demonstrate acts of altruistic behaviors that might be construed as benevolent and kind. The key, of course, lies in what and how the programmers design these systems, and what aspects of the human "Self" the AI might eventually mirror.

References

[1] Qi, C., & Lyu, J. (2022). Applications of artificial intelligence in children and elderly care and short video industries: Cases from CuboAI and Tiktok. International Conference on Computer Application and Information Security (ICCAIS 2021). https://doi.org/10.1117/12.2637376

[2] Rubeis, G. (2020). The disruptive power of artificial intelligence. ethical aspects of Gerontechnology in elderly care. Archives of Gerontology and Geriatrics, 91, 104186. https://doi.org/10.1016/j.archger.2020.104186

[3] Padhan S, Mohapatra A, Ramasamy SK, Agrawal S. Artificial Intelligence (AI) and Robotics in Elderly Healthcare: Enabling Independence and Quality of Life. Cureus. 2023 Aug 3;15(8):e42905. doi: 10.7759/cureus.42905. PMID: 37664381; PMCID: PMC10474924.

[4] Hasan, D. S., Pant, B., Kumar, Y., Rao, A., Singh, Y., & Srivastava, A. (2023b). Microrobot for elderly care using advance AI technology. 2023 IEEE World Conference on Applied Intelligence and Computing (AIC). https://doi.org/10.1109/aic57670.2023.10263815

[5] https://www.coursera.org/articles/types-of-ai

[6] Hintze, A. (2016)Understanding the four types of AI, from reactive robots to self-aware beings in *The Conversation.* https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616

[7]. Saikiran Chandha (2021) What if AI becomes self-aware? *Express Computer.* *https://www.expresscomputer.in/artificial-intelligence-ai/what-if-ai-becomes-self-aware/81828/*

[8] Gerardo, H., Egushi, A., Twomey, R., beyond the black box: four explorations in embodied interaction. *International Conference on Education, Research, and Innovation.* In press. 2023.

[9] Sparkes, M. (2023) AI shows no signs of consciousness yet, but we know what to look for. In *New Scientist.* August 30, 2023.

[10] Taylor, C. (1989) Sources of the Self. Cambridge, MA. Harvard University Press.

[11] Foucault, Michel. "The Subject and Power." In *Power. Essential Works of Foucault 1954-1984 Volume 3*. Edited by James D. Faubion, 326-348. New York: The New Press, 2000.

[12] Foucault, Michel. *Discipline and Punish*. Translated by Alan Sheridan. London: Penguin, 1977.

[13] Foucault, Michel. The History of Sexuality Vol. 1: An Introduction. Translated by Robert Hurley. New York: Vintage, 1990.

[14] Foucault, Michel. "Technologies of the Self." In *Ethics, Subjectivity, and Truth. The Essential Works of Michel Foucault Vol. 1*. Edited by Paul Rabinow. Translated by Robert Hurley et al., 223-251. New York: The New Press, 1997.

[15] Srinanth Srinivasa and Jayati Deshmukh, "AI and the Sense of Self". https://arxiv.org/abs/2201.05576 https://doi.org/10.48550/arXiv.2201.05576 Submitted on 7 January, 2022.

[16] Bryson, J., 2010. "Robots should be slaves." Wilks (Ed.), *Close Engagements with Artificial Companions*. John Benjamins.

[17] Metzinger, T., 2021. "Artificial suffering: An argument for a global moratorium on synthetic phenomenology." *Journal of Artificial Intelligence and Consciousness*, 8, pp.43–66.

18] Graziano, M. S., 2017. "The attention schema theory: a foundation for engineering artificial consciousness." *Frontiers in Robotics and AI*, 4(60).

[19] Schwitzgebel, E., & Garza, M., 2020. *Designing AI with rights, consciousness, self-respect, and freedom*. Liao (Ed.), Ethics of Artificial Intelligence. Oxford University Press.

[20] Weil, E., "You Are Not a Parrot: And a chatbot is not a human. And a linguist named

Emily M. Bender is very worried what will happen when we forget this." In *Artificial Intelligence.* March 1, 2023. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

[21] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). "Consciousness in artificial intelligence: Insights from the science of consciousness." *arXiv preprint arXiv:2308.08708*.

[22] Seglow Jonathan. *The Ethics of Altruism*. Dec 2022. DOI: 10.1080/13698230410001702702

[23] Clavien, C. and Chapuisat, M. (2013) "Altruism across Disciplines: One Word, Multiple Meanings" in *Biology & Philosophy 28(1):125–140*
DOI: 10.1007/s10539-012-9317-3

[24] Gallagher, J. "Fertility rate: 'Jaw-dropping' global crash in children being born" In *BBC.* https://www.bbc.com/news/health-53409521

[25] Amol Deshmukh, Srinivasan Janarthanam, Helen Hastie, Mei Yii Lim, Ruth Aylett, and Ginevra Castellano. 2016. How Expressiveness of a Robotic Tutor is Perceived by Children in a Learning Environment. In The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16). IEEE Press, Christchurch, New Zealand, 423–424.

[26] Niklas Möller, Sven Ove Hansson, Principles of engineering safety: Risk and uncertainty reduction, Reliability Engineering & System Safety, Volume 93, Issue 6, 2008, Pages 798-805, ISSN 0951-8320, https://doi.org/10.1016/j.ress.2007.03.031.