A Self for robots: core elements and ascription by humans

Sara Incao*
sara.incao@iit.it
Istituto Italiano di Tecnologia - CONTACT
University of Genoa

Francesco Rea
francesco.rea@iit.it
Istituto Italiano di Tecnologia - RBCS

Alessandra Sciutti
alessandra.sciutti@iit.it
Istituto Italiano di Tecnologia - CONTACT

ABSTRACT

Modern robotics is interested in developing humanoid robots with meta-cognitive capabilities in order to create systems that have the possibility of dealing efficiently with the presence of novel situations and unforeseen inputs. Given the relational nature of human beings, with a glimpse into the future of assistive robots, it seems relevant to start thinking about the nature of the interaction with such robots, increasingly human-like not only from the outside but also in terms of behavior. The question posed in this abstract concerns the possibility of ascribing the robot not only a mind but a more profound dimension: a Self.

KEYWORDS

Human-Robot Interaction, humanoid robots, artificial Self, Self

DOI

https://doi.org/10.5281/zenodo.5645583

Figure 1: Two iCub robots. One looking in a mirror and recognizing being different from the other iCub.

1 INTRODUCTION

If years ago, the nature of human artefacts called tools was clearly defined and those artefacts were perceived as objects built to facilitate a task, now that technology has evolved and those tools have gained human-like bodies and a certain degree of autonomy, the boundary between the status of object and subject turns to be blurred. Advancements in social robotics are moving towards the building of systems that are human-like not only from the point of view of physical appearance but also in terms of behavior during interactions [5, 14, 15].

Indeed, many scientists claim that all knowledge occurs in the background and basis of our corporeity since the very beginning of our life. This is the reason why a human-like body is such a fundamental element for a robot to have, if the goal is to become effective in assistance and collaboration with humans. If the future of assistive and homecare robots is that of being highly adaptive to human needs [26] and since adaptivity is mutual in a relationship, it is crucial to understand how humans might react to robots that show such a similar conduct to them. In other words, the question is what kind of attribution occurs when a human being is interacting with a robot, specifically, a humanoid robot that apparently perceives, reasons and acts as a human being.

Previous studies on this topic have mainly relied on Theory of Mind as a paradigm to investigate such query. Theory of Mind (ToM) [1] is the ability to attribute mental states to others and it has been found that humans use the same mechanism to attribute beliefs and intentions to robots by assigning them human-like traits or characteristics [6, 10, 28, 30]. Theory of Mind refers to the capacity to infer others' mental states: it is the ability of reasoning on the world, founded on the conviction that the representation of the world, rather than the world itself, determines others' actions. This metarepresentational capacity is fundamental when it comes to establish social relationships because understanding others' mental states are similar to ours, enables us to have meaningful social interactions [9, 25].

The concept of Intentional Stance [7, 8] is strictly linked to ToM and according to S. Marchesi et al. [16] it refers, as ToM, to the inference of others' mental states but regardless of the fact that the inference is correct. The failure of the classic Sally-Anne ToM test implies the inability to realize that other people have different mental states from ours: the inference about others' mental states is not simulated by analogy but by identification in the sense that we attribute our beliefs to the other person. In this case, it is the world as we see it that determines the other's actions. But, if we fail the simulation and therefore the Sally-Anne test giving the wrong answer, we are still convinced that Sally will look somewhere. Even though the inference in not correct, we will be still inferencing something about Sally's mental states and this means adopting the intentional stance.

Despite the different empirical ways in which ToM and Intentional stance are addressed, they have the common reference to the inference of mental states during interaction. Now, in the field of Human-Robot Interaction (HRI), posing the interrogative in terms of ToM is extremely important because it is necessary to understand and study how people tend to mentalize machines with a human-like appearance and behavior. However my proposal in this extended abstract intends to answer to a distinct question; I propose a new interrogative through which investigate the relationship that occurs when humans interact with robots: the ascription of a Self.

If ToM describes the way we make inferences about others' representation of the world attributing them mental states; the concept of Self is rather a definition of the entire horizon of experience of a person that takes place at a bodily dimension. Recognizing another "Self" means being in contact with another entity whose presence in the world we recognize -to some extent- as similar to ours. Rather than asking if we are likely to attribute mental states to robots, I propose to ask if we are likely to recognize other Selves in humanoid robots.

The ascription of a Self implies the identification of something underlying the temporary mental states [13]: there is a common ground that ties together the flow of experiences that accumulate in the life of a person which become their identity and distinctive feature. Indeed, the Self has been defined as an «active process of culturally mediated internalization of social interactions along multiple time scales» [2]. It is a fact that the people we know are present in our minds not just as entities with mental states but represented with specific features, first of all different bodies, but also different experiences and personalities constituting the Self. The question that now arises is whether we are likely to recognize a Self in a robot during an interaction with it.

2 THE SELF

The famous Nagel argument "What is it like to be a bat?" [18] shows that the subjective character of experience is something that defies any attempt to be precisely described. By common-sense, we can imagine that the experience of being a bat is entirely different from the experience of being a human. However, we will never be able to really feel what it is like to be a bat. From this argument it is clear that the subjective experience of all the living beings is not accessible from the outside.

Therefore, the discussion about the "Self", or many types of selves [21], derives from an operational abstraction of self-knowledge. I am a human being with a subjective experience that is not accessible to others because it is a first-person experience. The concept of Self is an operational definition aimed at giving the possibility of talking about something that has no defined features, nor a defined structure but it is what we are as living beings. All the subjective experiences, all the mental states, the bodily sensations, our own ways to relate to others and to the environment, our private emotions and everything related to our life and story constitutes the peculiar unity of one – more or less complex – living being.

3 ARTIFICIAL SELF FOR ROBOTS

«Social robots need a model of the "Self" » [15]: this claim by Mark H. Lee summarizes the most recent studies in the field of cognitive robotics. The question arises as to whether it is time for robots to become more social and less mechanical in view of a future in which the interaction between humans and robots will be a consistent practice. Since adaptivity and autonomy are both highly desirable features in HRI, several studies are focusing on the building of robots with an artificial Self. These features could also allow the system to react more accurately to the emotional, attentional and cognitive states of the observer [3, 22, 24].

In the field of robotics, the issue is considered in two respects. On the one hand, from an architectural approach, the aim is to build a cognitive architecture composed by several modules [11, 13, 20]. An open question is whether the Self should be a module among others or it ought to emerge from the complex connections of the architecture modules. On the other hand, the developmental approach is inspired by developmental studies in psychology. This approach is focused on the search for the emergence of some core elements of the Self, e.g.: self-other recognition, body ownership or sense of agency, etc., [2] through mechanisms of minimization of prediction error with recurrent neural networks in a predictive learning approach [25].

4 ASCRIBING A SELF TO A ROBOT IN HRI

Given these recent implementations, this paper proposes on the possibility of ascribing a Self to a robot, that is to say, the possibility of ascribing a complex set of distinctive features to it, similarly to those we are used to recognizing in people. In view of the objective pursued by developmental and cognitive robotics to provide robots with a sense of Self to enhance their adaptivity and autonomy in interactions with humans, it would be interesting to investigate how such a robot may be perceived by humans interacting with it. Thinking about a possible future scenario, a robot could be able to integrate features such as biological motion and multimodal perception (Spatial persistence of the Self), emotional alignment, joint attention and styles of action [27, 29] (Relational nature of the Self), adaptive behavior, face recognition and memory of past interactions [12, 26] (Temporal persistence of the Self) and the ability to recognize its own state and to predict the possible consequences of its own actions within the environment [23] (Metacognition).

To this aim, it is useful to define four components (see fig. 2) that can be used as operational definitions to address both the question about how to let a Self emerge in a robot and also, the one about the possibility of ascribing a Self to a robot.

4.1 The spatial persistence of the Self (inspired by Neisser's Ecological Self)

The Ecological Self defined by Neisser refers to the fact that we, as humans, are equipped with a particular visual apparatus and our view of the world can only take place from a particular point of view. Anything that moves together with the point of observation or anything that occurs in its visual field is perceived as part of the Self. This aspect of the self is present since the very beginning of our life because, as soon as we are brought into the world, we find ourselves placed in the environment. Even though the environment will certainly change, what will remain constant is the point of view from which we observe it. All the proprioceptive, kinesthetic and tactile sensations that reach our body provide the individual not only a placement in the space. They also provide a dynamic reference point within a space which becomes anthropologic in terms of being the space of our action in the world, the place in which our existence unfolds and is concretely felt. This is the reason why, ascribing a Self to a robot means recognizing in it this very disposition towards the surrounding environment. Its peculiar attitude towards the world becomes the distinctive feature that may lead to identify a Self in it.

4.2 The relational nature of the Self (inspired by Neisser's Interpersonal Self)

The Interpersonal Self that Neisser describes is the engagement in a social interaction. Murray and Trevarthen [17] found that infants from 6 to 8 weeks old are able to distinguish the experience of a real time interaction in which the other person is reciprocally engaged with them, from the mere view of a recording of that interaction in which they don't feel the real time engagement. Since birth we are part of a social environment in which the presence of other people is constant, but the development of the interpersonal Self is due to the reciprocity of the exchange with others. Even though the concept of interaction includes also a collaborative task performed by a human and a robotic arm, when it comes to design experiments aimed at recognizing an Interpersonal Self in a robot, the element of reciprocity must be considered. Thinking about a concrete experimental scenario, it could be therefore interesting to verify whether and under which circumstances adaptive or not adaptive robots are considered to have a Self by eliciting different degrees of engagement in the interaction. Moreover, another question that might arise is whether the human-like or not human-like body of robots with the same adaptive behavior, affects human participants' engagement.

4.3 The temporal persistence of the Self

The distinction of one from the other among humans is grounded on the recognition of an individual unity, a precise individuality that cannot be modified despite the infinite number of adjustments that one person undergoes over lifetime. This is the reason why the temporal extension of a relationship is critically important to be considered when the question is about the Self. All the individual features through which the representation of other people takes form for us are strictly related to temporality. When we interact with someone that we have already met before, we keep in mind the representation of this person that has evolved over the past encounters and, on the basis of this peculiar image of their Self, we carry on the conversation. Since the time factor is so crucial in human-human interactions, I suggest for HRI to consider the fact of designing extended in time experimental scenarios with repeated interactions with the same robot to explore to which extent the image of a robot formed over time may induce the subject to recognize in it a Self or not. Another interesting question that could be addressed with two robots of the same type, both appearing and behaving exactly the same, is whether humans tend to standardize the Self, the identity of one robot, by transferring it to the other.

4.4 Metacognition [4] (inspired by Neisser's Conceptual Self)

In humans, introspection is a complex process that allows reflection on the self. The network of assumptions and beliefs about ourselves is fundamental to give rise to a unified image of who we are that constitutes our reference point for ourselves and is coherent over time. While at birth, in the ecological and interpersonal selves, the relationship self-world and self-other is direct, immediate, without the conceptualization of categories such as "the Self", "the world" and "the other", the development of metareasoning allows the abstraction and therefore the conceptualization of what was directly perceived before. Consequently, although metareasoning on the Self develops at the end of the first year of life, [19] it is not something that appears from nothingness at a certain point in growth: its prerequisites are present since the first day of our life. Even though they are not expressed because they lack the possibility of abstraction, a direct and unmediated reference to the Self is always present. As a matter of fact, when newborns move their limbs, the direct perception of the environment occurs in relation to their limbs, to their body that is, for now, only the ground zero of every perception. Gradually, this ground zero begins a process of abstraction in which it acquires a shape and, in the end, through metareasoning, it can be taken as object of thought.

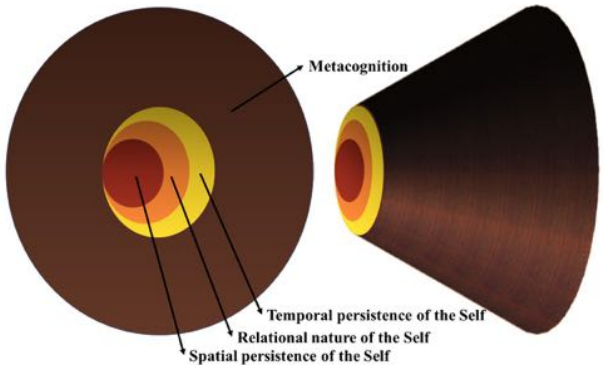

Figure 2: Representation of the Self through its four components. All of them emerge and develop from the same core. Metacognition unfolds at a certain point in life allowing reflection on all the other aspects of Self

5 CONCLUSION

Our experience is much more than a simple reaction to something outside, it concerns the way we see the world and the internal representations of it. It is not at all easy to reduce the complexity of a human Self to categories and elements but it has to be considered in the context of an attempt to build a descriptive, not exhaustive, model of our ability to interact effectively and significantly with the physical and social environment. In the field of Human-Robot Interaction, since the recent studies are addressed towards the implementation of meta-cognitive capabilities, understanding if humans are likely to recognize the depth of a Self in robots is fundamental to decide which direction to take and whether there are limitations to the complexity we can ascribe to humanoid machines.

ACKNOWLEDGMENTS

We thank Professor Giulio Sandini for his useful suggestions. A.S. was supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. G.A. No 804388, wHiSPER

REFERENCES

[1] Simon Baron-Cohen. 1995. *Mindblindness: An Essay on Autism and Theory of Mind.* The MIT Press.
[2] Dimitris Bolis and Leonhard Schilbach. 2018. 'I Interact Therefore I Am': The Self as a Historical Product of Dialectical Attunement. *Topoi* 39, 3 (2018), 1–14. https://doi.org/10.1007/s11245-018-9574-0

[3] Ronald J. Brachman. 2002. Systems That Know What They're Doing. *IEEE Intelligent Systems* 17, 6 (2002), 67–71. https://doi.org/10.1109/MIS.2002.1134363

[4] Cristiano Castelfranchi and Rino Falcone. 2018. Self-Awareness implied in human and robot intentional action. *CEUR Workshop Proceedings* 2287, 6 (2018), 1–7.

[5] Raja Chatila, Erwan Renaudo, Mihai Andries, Ricardo Omar Chavez-Garcia, Pierre Luce-Vayrac, Raphael Gottstein, Rachid Alami, Aurélie Clodic, Sandra Devin, Benoît Girard, and Mehdi Khamassi. 2018. Toward self-aware robots. *Frontiers Robotics AI* 5, AUG (2018). https://doi.org/10.3389/frobt.2018.00088

[6] Maartje MA De Graaf and Bertram F. Malle. 2019. People 's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2019), 239–248.

[7] Daniel C. Dennett. 1971. Intentional systems. *The Journal of Philosophy* 68, 4 (1971), 87–106.

[8] Daniel C. Dennett. 1989. *The Intentional Stance.* MIT Press, Cambridge, MA.

[9] Christine Evans-Pughe. 2013. *Mapping the mind.* Vol. 8. 42–44 pages. https://doi.org/10.1049/et.2013.0202

[10] Friederike Eyssel, Dieta Kuchenbrandt, Frank Hegel, and Laura De Ruiter. 2012. Activating elicited agent knowledge: How robot and user features shape the perception of social robots. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (2012), 851–857. https://doi.org/10.1109/ROMAN.2012.6343858

[11] Shaun Gallagher. 2000. Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences* 4, 1 (2000), 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5

[12] Jonas Gonzalez-Billandon, Alessandra Sciutti, Matthew Tata, Giulio Sandini, and Francesco Rea. 2020. Audiovisual cognitive architecture for autonomous learning of face localisation by a Humanoid Robot. *Proceedings - IEEE International Conference on Robotics and Automation* (2020), 5979–5985. https://doi.org/10.1109/ICRA40945.2020.9196829

[13] Michael Graziano and Taylor W Webb. 2018. Understanding Consciousness by Building It. In *The Bloomsbury Companion to the Philosophy of Consciousness*, Dale Jacquette (Ed.).

[14] Verena V. Hafner, Pontus Loviken, Antonio Pico Villalpando, and Guido Schillaci. 2020. Prerequisites for an Artificial Self. *Frontiers in Neurorobotics* 14, February (2020), 1–10. https://doi.org/10.3389/fnbot.2020.00005

[15] Mark Lee. 2020. *How to Grow a Robot.* MIT Press, Cambridge, Massachussets.

[16] Serena Marchesi, Davide Ghiglino, Francesca Ciardo, Jairo Perez-Osorio, Ebru Baykara, and Agnieszka Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology* 10, MAR (2019), 1–13. https://doi.org/10.3389/fpsyg.2019.00450

[17] Lynne Murray and Colwyn B. Trevarthen. 1985. Emotional Regulation of interactions between 2 months old and their mothers. In *Social perception in infants.*

[18] Thomas Nagel. 1974. What Is It Like to Be a Bat? *The philosophical review* (1974).

[19] Ulric Neisser. 1994. *The Perceived Self: Ecological and Interpersonal Sources of Self Knowledge.*

[20] Ulric Neisser. 1995. *Criteria for an ecological self.* Vol. 112. Elsevier Masson SAS. 17–34 pages. https://doi.org/10.1016/S0166-4115(05)80004-4

[21] Ulric Neisser. 2008. Five kinds of self - knowledge. *Philosophical Psychology* June 2013 (2008), 37–41.

[22] Rony Novianto and Mary Anne Williams. 2009. The role of attention in robot self-awareness. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (2009), 1047–1053. https://doi.org/10.1109/ROMAN.2009.5326155

[23] Armin Sadighi, Bryan Donyanavard, Thawra Kadeed, Kasra Moazzemi, Tiago Muck, Ahmed Nassar, Amir M. Rahmani, Thomas Wild, Nikil Dutt, Rolf Ernst, Andreas Herkersdorf, and Fadi Kurdahi. 2018. Design methodologies for enabling self-awareness in autonomous systems. *Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition, DATE 2018* (2018), 1532–1537. https://doi.org/10.23919/DATE.2018.8342259

[24] Brian Scassellati. 2002. Theory of Mind for a Humanoid Robot Approved for Public Release Distribution Unlimited. *Autonomous Robots* 12, 1 (2002), 13–24.

[25] Brian J. Scholl and Alan M. Leslie. 1999. Modularity, development and 'theory of mind'. *Mind and Language* 14, 1 (1999), 131–153. https://doi.org/10.1111/1468-0017.00106

[26] Ana Tanevska, Francesco Rea, Giulio Sandini, Lola Cañamero, and Alessandra Sciutti. 2020. A Socially Adaptable Framework for Human-Robot Interaction. *Frontiers in Robotics and AI* 7, October (2020), 1–16. https://doi.org/10.3389/frobt.2020.00121

[27] Ana Tanevska, Francesco Rea, Giulio Sandini, and Alessandra Scuitti. 2017. Can emotions enhance the robot's cognitive abilities: A study in autonomous HRI with an emotional robot. *Proceedings of AISB Annual Convention 2017* April (2017), 204–208.

[28] Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology* 8, NOV (2017), 1–14. https://doi.org/10.3389/fpsyg.2017.01962

[29] Fabio Vannucci, Giuseppe Di Cesare, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. 2019. A Robot with Style: Can Robotic Attitudes Influence Human Actions? *IEEE-RAS International Conference on Humanoid Robots* 2018-Novem (2019), 952–957. https://doi.org/10.1109/HUMANOIDS.2018.8625004

[30] Robert H. Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems. *IJCAI-2016 Ethics for Artificial Intelligence Workshop* Dc (2016). http://opus.bath.ac.uk/50294/1/WorthamTheodorouBryson{_}EFAI16.pdf