From Identified to Self-Identifying: Social Identity Theory for Socially Embodied Artificial Agents

Katie Seaborn
seaborn.k.aa@m.titech.ac.jp
Tokyo Institute of Technology
Tokyo, Tokyo, Japan

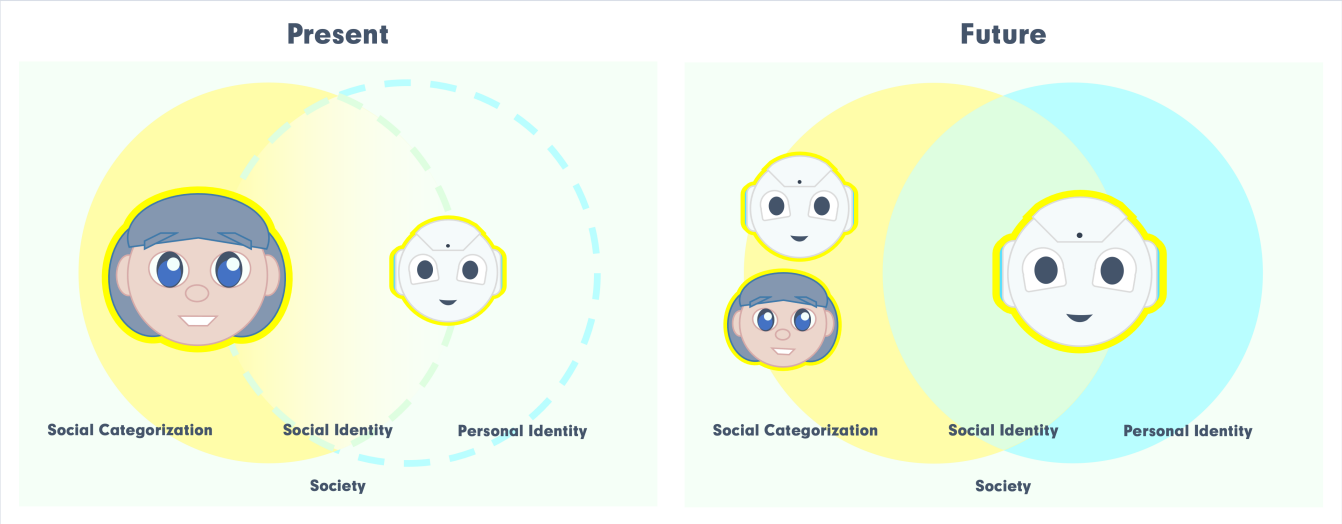

Figure 1: Two perspectives on robot identity. Based on social identity theory.

ABSTRACT

Robots and other artificial agents are growing in sophistication and sociability, leading many to consider higher-order social factors like identity. In this position paper, I present two perspectives on robot identity for the present and the future by extending the classic human model of *social identity theory*. I propose a way of determining the degree to which robots, as autonomous entities embedded in human societies, may achieve the ability to participate in social identity activities. I argue that we are a long way off from robot self-identity. Nevertheless, we should decide what our criteria are now with a firmly ethical stance and scholarly rigour.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design theory, concepts and paradigms**; • **Computer systems organization** → *Robotic autonomy*; • **Computing methodologies** → *Cognitive robotics*; • **Applied computing** → *Sociology*.

KEYWORDS

robots, social robots, user perception, identity, social identity theory, self-identity, autonomy, social perception, social embodiment

1 INTRODUCTION

Advances in social robotics have begun to inspire new questions and avenues for research on higher-order social characteristics, attitudes, and behaviours. One such factor is *identity*, specifically whether and how robots and other artificial agents can have an identity. The notion of "robo-identities" or "artificial identities," to be inclusive of non-robots, remains understudied, and perhaps for good reason. A range of human endeavours, notably science fiction in every conceivable medium, have provided visions of robots and other artificial agents navigating issues of identity. Martha Wells's Hugo- and Nebula-award winning *The Murderbot Diaries* (2017-) is a recent example. The series features the first-person perspective of self-anointed "Murderbot," a genderless hodgepodge of biology and machine who comes into self-awareness and the ability to self-identify. While compelling, such visions are far flung in time. Artificial agents that have advanced to this extent do not yet exist. But we can concertedly and conscientiously aim in this direction.

In this position paper, I nominate a classic model of human identity, *social identity theory*, as a germinating framework. Social identity theory explains the relationship between the (human) self and one's social world, including how (human) others influence one's own self-identity, as well as our propensity to perceive identities in others and/or assign identities to others. Here, I present a version of this theory that is inclusive to robots and other artificial agents. I show how it may then be used to characterize the degree to which robots have and have not achieved the ability to engage in social identity activities, in the present and in the future. I argue that we are in the early stages but we can, and should, prepare

ourselves with a rubric now, to direct research and development as well as to avoid future ethical quandaries.

2 SOCIAL IDENTITY THEORY AND ARTIFICIAL AGENTS

Social identity theory [3, 12–14] was developed over decades by Tafejl, Turner, Hogg, and others to explain how the social intersects with the individual. Specifically, this theory considers how we conceptualize the self in relation to others, as well as how our self-conception intersects with the social groups to which we identify as belonging. In this way, it offers an interactionist perspective that links the attitudes and behaviours of the self and with one's social group memberships. While originally formalized over forty years ago, it has stood the test of time, remaining well-employed across many fields of study to explore the self in intergroup relations. *Social constructionism* [1] is at the heart of the theory: people, and potentially all agents of certain social intelligence, are embedded in a social world, i.e., society, and thereby fundamentally influenced by it when it comes to the development of identity. Societies are, of course, made up of other social actors who may ascribe identities onto others. The plurality of "identity" is key: multiple identities are likely, if not assured [4]. In a nutshell, the theory maps out three intertwined components of identity formation and ascription:

- *Personal Identity*: How we define ourselves individually.
- *Social Identity*: How we define ourselves in relation to others, i.e., social group membership.
- *Social Categorization*: How others define us, and how we define others, i.e., social perceptions.

These components are acted upon in three key ways, which I have framed around the case of a generic "agent" so as to include not only humans but potentially robots and other artificial beings:

- *Social Categorization*: Agents categorize others, i.e., *being identified, others' perceptions of the self*).
- *Social Identification*: Agents categorize themselves, i.e., *self-identifying, one's own perceptions of the self*).
- *Social Comparison*: Agents compare themselves to others based on the group with which they identify, i.e., *in-group and out-group identification*).

While a useful framework for describing people, social identity theory may also be extended to non-humans. Indeed, we can interpret these fundamental pieces—each set of components and actions—as a rubric for measuring the extent to which any agent has attained the ability to participate in identity activities. A robot may be able to classify you as Canadian or young or irreligious. It may perceive itself as of a mechanical gender or liberal or short. It may recognize that it was created in Japan based on subtle features in its visual appearance, even though it may have a British nationality on account of its present living (generously stated) conditions and the accent it has subsequently picked up—all this by comparison to the social agents with whom it interacts and the social identities within its roster. It may reject human models and devise, in concert with other selfsame robots, its own.

Explorations of social identity theory and artificial agents pepper the last couple of decades, understudied and in limited form. Notably, most work has taken a one-sided perspective, out of necessity: at present, only humans have the capacity to exercise the full breadth of the theory. Social embodiment remains a fuzzy and unstable state, Most recently, Edwards et al. [2] explored student identification with an AI instructional agent, finding that age identification was critical. They also persuasively argued that social identity theory is premised in the Computers Are Social Actors (CASA) model originally developed by Nass, Moon, Lee, Brave, and colleagues [5–7]. According to the CASA, people reflexively apply human models to artificial agents, often without realizing it and usually in accordance with dominant human stereotypes. A recent survey of work on voice in human-agent interaction (vHAI) by Seaborn and colleagues revealed a temporal, and possibly generational, component to this [9]. Mixed results from a rapid review of gender neutrality in human-robot interaction (HRI) [10] combined with new evidence of researcher influence on participant gendering in the case of humanoid robot Pepper [8] also point to the same underlying, unidirectional phenomenon. What is missing is the bidirectional, truly social element.

I have illustrated this state of affairs in 1. On the left, we have the present situation: people, rather than the agent itself, decides on the agent's identity. The gradient within the overlapping bounds of the social identity component points to the influence of people on the robot's identity according to how it has been designed and how that design is interpreted in line with particular human social groups. For instance, someone may interpret a robot as having their own nationality and express greater warmth towards it on that basis alone [11]. On the right, we have the future, which represents what people are able to do now. The agent will have its own internal identity, which is premised in social identities, which are recognized by other agents, including (but not necessarily) people. Attaining this level of social identity ability will require a level of social intelligence and ability to interface with the world that does not yet exist. Such agents will need to pick up on and adapt to the norms of various social identities: a multi-sensory endeavour that would recognize and take part in the shifting and evaluative nature of social identities. Indeed, an agent would need to internalize but also influence the social categories that it has decided to adopt.

3 IMPLICATIONS

I will now propose some ways in which achieving the quest for social identity in artificial agents may affect us.

- *Identity Multiplicity*: Agents may carry multiple and likely intersecting identities at the same time.
- *Identity Rejection*: Agents may reject the categories imposed by others, including people, such as by shirking role assignments and questioning stereotyped reactions.
- *Identity Favouritism*: Agents may prefer some people over others. Agents may prefer non-humans over humans.
- *Identity Creativity*: Agents may have unusual and creative reasons for perceiving themselves as part of one group and not the other, challenging convention and the status quo.
- *Identity Unpredictability*: Future encounters with the same agent may not proceed in line with expectations due to shifts in personal identity or the norms of the social identities with which the agent identifies.

- *Identity Mobility*: Similar to the above, agents may adopt a social mobility strategy and shift groups based on the perceived social value of all available groups.
- *Identity Influence*: Agents may influence human social identities. Full participation is a two-way street.
- *Identity Sacrifice*: Agents may be willing to sacrifice on behalf of in-group members and not others.
- *Identity Passing*: Agents may "pass" as a member of a group despite not meeting all of the criteria that define that group. In particular, agents may pass as human when they are not.

This is by no means a complete list. Rather, it is a curated selection based on my understanding of social identity theory and my expertise is human-agent interaction (HAI) and HRI. We cannot know whether and how these implications may play out. Human models may not transfer to the agents of the future, despite being crafted by human hands. I have also focused on the theory at a high level without accounting for the plethora of subtheories representing the nitty-gritty of reality, e.g., self-categorization theory [14]. Particular subtheories may be more germane than others, now and in the future. We humans are also still working towards understanding ourselves, and so we will need to return to theorizing this extension when we update the human one.

4 CONCLUSIONS

New opportunities for robot identity are on the horizon. We can prepare for the socially intelligent future with our artificial companions now. Social identity theory is but one model to consider. In the end, what direction our mechanical confederates take may be out of our hands. And, if this comes about, it may signal our success in crafting a form of life that is truly self-identifying. Whether or not we wish to enable this future is a question for another day.

ACKNOWLEDGMENTS

Thank you to Peter Pennefather for ongoing inspiration through our discussions about artificial agents and identity.

REFERENCES

[1] Vivien Burr. 2015. *Social Constructionism.* Routledge.
[2] Chad Edwards, Autumn Edwards, Brett Stoll, Xialing Lin, and Noelle Massey. 2019. Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior* 90 (2019), 357–362.
[3] Michael A Hogg. 2016. Social identity theory. In *Understanding peace and conflict through social identity theory.* Springer, 3–17.
[4] Michael A Hogg, Deborah J Terry, and Katherine M White. 1995. A tale of two theories: A critical comparison of identity theory with social identity theory. *Social psychology quarterly* (1995), 255–269.
[5] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI'00 extended abstracts on Human factors in computing systems.* 289–290.
[6] Clifford Nass, Youngme Moon, Brian J Fogg, Byron Reeves, and D Christopher Dryer. 1995. Can computer personalities be human personalities? *International Journal of Human-Computer Studies* 43, 2 (1995), 223–239.
[7] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press Cambridge.
[8] Katie Seaborn and Alexa Frank. 2022. What pronouns for Pepper? A critical review of gender/ing in research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* ACM, New Orleans, Louisiana, USA, accepted. https://doi.org/10.1145/3491102.3501996
[9] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–43.
[10] Katie Seaborn and Peter Pennefather. 2022. Neither "hear" nor "their": Interrogating gender neutrality in robots. In *Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction.* ACM/IEEE, Sapporo, Hokkaido, Japan, accepted.
[11] Nicolas Spatola, Nolwenn Anier, Sandrine Redersdorff, Ludovic Ferrand, Clément Belletier, Alice Normand, and Pascal Huguet. 2019. National stereotypes and robots' perception: the "made in" effect. *Frontiers in Robotics and AI* 6 (2019), 21.
[12] Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior. In *Political psychology.* Psychology Press, 276–293.
[13] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56, 65 (1979), 9780203505984–16.
[14] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. 1987. *Rediscovering the social group: A self-categorization theory.* basil Blackwell.