Design, Performance, and Perception of Robot Identity

Ryan Blake Jackson
rbjackso@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Alexandra Bejarano
abejarano@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Katie Winkle
winkle@kth.se
KTH
Stockholm, Sweden

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, Colorado, USA

ABSTRACT

This paper explores (1) how robots and multi-robot systems can perform identity specifically for human benefit, (2) the factors that impact how humans perceive robot identity and its connection with mind and body, and (3) possible implications of designing non-traditional identity configurations. In particular, we explore the unique ways that identity may be performed in multi-robot systems, and examine arguments for and against designing multi-robot systems to perform identity in ways that diverge from or obscure the distributed nature of those robots' cognitive architecture.

1 INTRODUCTION

Mind, body, and identity in humans are typically understood as having a default 1-1-1 mapping such that every human has a single unique and inherent mind, body, and identity. While some philosophers have begun to explore the ways that these constraints might be broken by, for example, offloading cognitive functioning onto cognitive technologies [5], the majority of explorations beyond a 1-1-1 connection of mind, body, and identity have remained speculative (see, e.g., [2]). Robots, however, are not bound by this traditional mapping, and there are many well-known strategies for organizing robot minds, bodies, and identities beyond simple 1-1-1 correspondence, especially in multi-robot systems composed of multiple robotic bodies, minds, and/or identities.

Researchers have explored a number of alternative configurations in recent years. Luria et al. [13], for example, explores configurations involving one centralized robot mind/identity controlling multiple robot bodies (one-for-all), one robot body housing multiple distinct minds/identities (co-embodiment), and one mind/identity hopping between several bodies (re-embodiment). These schema can apply separately to both robot mind (i.e., whatever system is performing the robot's cognitive computation) and robot identity (which may be intuitively thought of as the individual, self, or persona perceived by others, though as we will see, a precise definition requires a more careful analysis). Critically, while mind and identity maintain a 1-1 association in humans, we view identity in robotics as purely performative, and thus decoupled from the robot's mind. Therefore, a multi-robot system could re-embody a robot mind, by changing which hardware is running the robot's software, without changing its performance of robot identity. Conversely, and perhaps more interestingly, a multi-robot system could performatively re-embody a robot identity without actually changing the loci of the constituent robots' cognition (minds). This paper explores several questions and potentials arising from this flexibility.

These performative changes of identity configurations, such as performative re-embodiment, appear differently from the user's perspective (in which alignment of mind, body, and identity appear to change) versus the developer's perspective (in which this alignment may not be viewed as truly changing). This means that identity can and must be analyzed differently from different perspectives. Like most HRI research, we are primarily concerned with how robot design choices impact the quality and nature of interaction with users, and we will therefore conduct our analysis of identity with the user's perspective in mind, viewing human interactants as *constitutive others*. That is to say, we view robot identities as existing, and configurations of identity as changing, insofar as these identities and configurations are perceived as existing and changing. Thus, before we can analyze how robot designers might create different identities and identity configurations in the minds of users, we will first formalize our focus on identity as perceived by users, and explain the different *observables* that may lead users to differentially perceive identities and identity configurations at their *level of abstraction*.

2 LEVELS OF ABSTRACTION

Several scholars have argued that it is necessary to clearly specify a *level of abstraction* (LoA) before discussing concepts such as robot (moral/social) agency [7]. A LoA consists of a collection of observables, each with a well-defined set of possible values or outcomes [6]. These observables determine how an entity may be regarded or described at different LoAs. Failure to specify a LoA invites inconsistencies, ambiguities, and disagreements stemming from unspoken differences in LoA, while specifying a LoA clarifies the range of questions that can be meaningfully asked and answered about a system, and the kind and amount of information that can be known regarding the system.

Floridi [6] presents wine as an example: a wine taster's LoA might consist of observables for sweetness, acidity, and tannicity, whereas a wine purchaser's LoA might consist of observables for price, maker, and vintage. These different observables denote what information is relevant and available from different perspectives.

As with other concepts in HRI like robot agency, a LoA must be specified before discussing robot identity. If, as described above, our conception of identity is based largely on human ascription/perception thereof, then identity is contingent on the perspective (or LoA) of whoever is doing the perceiving and ascribing, with identity only "existing" within the mind of an observer based on what they can perceive at their LoA. Moreover, identity must be regarded differently at different LoAs because different identity-relevant features are observable for different perceivers. Most HRI research is primarily concerned with the user's LoA, due to the impact users' perceptions and expectations regarding a robot can have on the efficacy of interactions. Indeed, most HRI research and design can be framed as manipulation of the observables that constitute the user's LoA to evoke desirable user perceptions and beliefs.

Observables relevant to identity at the user's LoA include external features of the robot. *Naming* is particularly salient for identity [3] as it frames the robot as a cohesive and unambiguous referent, and triggers cognitive processes in humans, such as reference resolution, that may create and reify mental representations of the robot's identity [18]. Moreover, naming can lead interactants to regard robots as *social* entities with mental states and emotions [1]. Robot *speech* and *behavior* are also important to constructing identity. Researchers have attached different voices to different robot identities to differentiate these identities as they migrate across bodies [13]. Others have enabled robot bodies in a one-for-all "hive mind" cognitive architecture to perform traditional 1-1 identity by performatively communicating among themselves verbally when humans are around, even though speech is not actually how they share information [17]. Such robot bodies could combine naming and speech observables by referring to each other by unique names, even though (at the developer's LoA) only one mind exists between them. Physical behaviors can also differentiate robot identities. For instance, by moving at different speeds, displaying different proxemic behavior, and choosing differently between options, robot bodies can give a sense of individual identity within a group [8]. Finally, robot bodies can provide visual identity cues, like a digital face displayed on a screen or a specific color displayed on LEDs, that could move with a robot identity between bodies [12, 13]. Heterogeneity across multiple robot bodies can also give a sense of individual identity within a group [8].

In contrast, the robot developer's LoA includes internal information, such as the mechanisms by which the robot perceives the world, represents knowledge, and selects actions. In a multi-robot system, the developer knows the true alignment of robot minds to bodies, whereas users must make inferences about this alignment from the system's behavior.

3 DESIGNING IDENTITY PERFORMANCE

We have discussed the observables that lead users to form impressions of robot identity. To users, such identities may seem inherent or emergent, as with human identity. However, to developers, robot identity performance is a flexible design choice to deliberately shape user impressions, and any given system could be made to perform various identities without changing its cognitive architecture. In this section, we explain *why* developers may leverage identity observables to make a robotic system perform an identity to users that does not correspond to the organization of minds and bodies at the developer's LoA (e.g., making each body perform a unique identity in a one-for-all cognitive system).

Above, we cited the example of robot bodies performatively verbalizing information to keep human teammates in the loop and at ease, despite a one-for-all organization of minds to bodies making such verbalization extraneous for information transfer. In general, it is easy to imagine that users might be most comfortable communicating with one-for-one identities in robots because that is what they are used to from human-human interaction, and research has found evidence for user discomfort during co-embodiment interactions [13]. This suggests that designers may wish to avoid observables suggesting co-embodiment in identity performance, even if co-embodiment is actually occurring in that the hardware of a single robot body is running multiple distinct artificial social actors[1]. However, the same study indicated positive reactions to identity re-embodiment in HRI, suggesting robot designers need not simply recreate humanlike configurations of identity.

Besides human comfort and ease of communication, researchers have cited trust as a critical consideration in designing robot identity performance [15, 16]. Specifically, a new theory of human-robot trust called *deconstructed trustee theory* treats robot identity and robot body as two distinct loci of trust, allowing the level of trust placed in each to differ. The first experimental evidence supporting deconstructed trustee theory indicates differences in human trust ascription depending on robot identity performance, with body-identity dissociating communication policies (i.e., performing non-one-for-one identity) decreasing the potential for robot bodies to be viewed as loci for capability trust [15]. We view deconstructed trustee theory as indicative of a potential broader paradigm wherein judgments of robots (e.g., trust, likeability, and attachment) are made separately for robot identities versus bodies.

Identity performance in a multi-robot system could also change based on context. Such "identity fluidity" in multi-robot systems can go beyond what an individual human can change about their identity (e.g., spinning up a new identity at will or switching which identities are in which bodies), which could be useful, for example, if a robot needs to do something that it predicts human teammates will dislike. We can imagine a one-for-all cognitive system where each robot body performs a unique identity (one-for-one) *and* the centralized mind of the system has an identity that can interact with users via co-embodiment of the many robot bodies it is controlling. In such a situation, a robot body could frame its undesirable actions in terms of belonging to a scapegoat identity, thus preserving esteem in the broader system.

For instance, the ability to refuse certain commands (e.g., immoral or infeasible ones) is important for morally and socially competent robots, but research on robot noncompliance interactions has indicated that refusing a human's command can damage robot likeability if the politeness of the refusal is miscalibrated to the context [11]. In such a situation, a robot might be able to frame its refusal as belonging to a narrower individual identity, insulating the broader system from any resultant damage to likeability. On the other hand, a moral rebuke from the centralized mind's identity might be more authoritative and persuasive to the human listener than a rebuke from a "less important" identity. Likewise, if a robot body does something praiseworthy, it could frame that accomplishment in terms of the larger centralized identity, and perform co-embodiment with that identity in reporting its accomplishment, but this may be less important given recent findings that human ascriptions of blame tend to be more intense and more subtly differentiated than praise [9, 15]. Empirical work will be necessary to determine the efficacy of such an approach.

[1]Consider a team of an autonomous land vehicle and a flying quadcopter. The land robot does not have weight constraints, so it can carry a lot of computing hardware. It might make sense to have both social actors running on the land robot's hardware so that the quadcopter can carry less computing hardware and be lighter.

4 COMPOSING AND FRAGMENTING IDENTITY

The above example of a system with one-for-all cognition that performs both one-for-one identities of each of its bodies and the larger centralized identity through co-embodiment raises the idea of hierarchically nested composite identities. Such composite identities exist to some extent in human systems. Corporations, research labs, and governments, for example, all have behaviors, goals, beliefs, unique names, and other identity cues. Despite being comprised of individual humans each with their own identities, we argue that the identity of the composite system is not simply some summation, combination, or average of the identities of its constituent humans. This is especially clear when these composite entities (particularly corporations) act in their own self-interest against the best interests of most of their constituent humans. However, composite identity in multi-robot systems differs in its interactional capabilities; one could have a conversation with the one-for-all robot mind, but one cannot have a conversation with a corporation (only a figurehead representing the corporation). Nonetheless, familiarity with other types of composite identities may help users to conceptualize multi-robot systems as composite identities.

Research on robot group entitativity touches on a type of implicit composite identity similar to that of informal human social groups. Highly entitative (i.e., more homogeneous and cohesive, which we argue gives a stronger impression of composite group identity) groups of robot bodies are perceived as more threatening than single robot bodies or diverse (less entitative, likely weaker sense of composite identity) groups of robot bodies [8]. Given results like these, robot designers may want to manipulate the observables governing group entitativity to make their multi-robot systems more palatable to human interactants. Even in systems where a group of bodies is centrally controlled by a single mind, making the group highly unified at the developer's LoA, straightforward variations between bodies in morphology, movement speed, pathing, and other simple behaviors can weaken impressions of entitativity, prompt impressions of individual identity within each body, and make the system's composite identity less threatening.

Having considered the building of composite identities from other sub-identities, we now consider the converse concept of fragmenting individual identities into their distinct constituent components. Just as the cognitive architectures community can blur the boundaries of what constitutes an individual mind via component sharing in multi-robot systems (e.g., by having two robots share dialogue systems while keeping other cognitive capacities separate) [14], so too can we break the rules of what normally constitutes or derives from identity in humans. For example, in robots there is not necessarily a static 1-1 mapping between identity and memory or identity and perceptual locus.

Breaking these conventions around identity raises many questions. For instance, previous researchers have wondered whether cognitive component sharing ties the robots' identities to their respective bodies more so than to their (partially) shared mind(s) [14]. However, knowledge of whether/how software components are shared is not directly available at the user's LoA, so we are less concerned with this particular question than with whether we should allow robot behaviors that evince component sharing observably (e.g., a robot remembering something that happened to a separate robot, or reporting what a separate robot is seeing). This divisibility of minds and identities also raises questions about what components of identity are necessary and constitutive versus simply coincidentally associated with identity because they co-occur in humans. For example, it is not clear whether beliefs, desires, and intentions are necessary parts of an identity or simply co-occur with human identity.

5 BREAKING THE ILLUSION

Throughout this paper, we have discussed deliberately shaping user perceptions of robots and multi-robot systems by performing identity in ways that obscure the actual implementations of such systems. This practice raises the question of how users might react to finding out that a system's identity was performed solely to influence their perception and, in that sense, was an illusion.

Users might perceive a system's identity performance as deceptive if an observable is added to their LoA that reveals the identity as purely performance, but robot deception is not always bad. Scholars have argued that robots actually ought to deceive people in certain ways, like benign prosocial "bullshitting" to ingratiate robots with interactants, despite the human intuition that there is something undesirable about being deceived [10, 20] (although cp. [19]). But despite the utility of certain types of robotic deception (like identity performance), human reactions may not be positive. Studies have shown that even relatively subtle deception in robot motion planning (making it seem like the robot intended to grab one object and then grabbing another) in an inconsequential game context where such deception was within the rules caused human interactants to trust the robot less because *it revealed the robot's capacity to deceive* [4]. Revealing an apparently deceptive identity performance could lead to a similar impression of the capacity for further, less benign deception and a corresponding drop in trust.

On the other hand, the same study showed some potentially positive effects of overt robot deception (beyond what we discussed in previous sections) in that the deceptive robot was rated as more intelligent, more engaging, and a better adversary in the game, and more so when the deception was perceived as intentional (though some of these effects were not significant from the study's 12 participants) [4]. Further experimentation is needed to ascertain whether similar positive impressions might come from revealing deceptive identity performance, and to generally determine what user reactions will be. Perhaps most importantly, we need to determine whether robot identity performance will still be effective if users view it as illusory or deceptive.

ACKNOWLEDGMENTS

This work was funded in part by a NASA Early Career Faculty Award and in part by an AFOSR Young Investigator Award.
Ryan Blake Jackson is a PhD student at Colorado School of Mines.
Alexandra Bejarano is a PhD student at Colorado School of Mines.
Dr. Katie Winkle is a Postdoctoral Research Fellow at KTH.
Dr. Tom Williams is an Assistant Professor of Computer Science at Colorado School of Mines.

REFERENCES

[1] Kate Darling. 2015. 'Who's Johnny?'Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy. In *ROBOT ETHICS 2.0*.

[2] Daniel C. Dennett. 1978. *Brainstorms*. Bradford Books, Chapter "Where Am I?".

[3] Kenneth L Dion. 1983. Names, identity, and self. *Names* 31, 4 (1983), 245–257.

[4] Anca D Dragan, Rachel M Holladay, and Siddhartha S Srinivasa. 2014. An Analysis of Deceptive Robot Motion.. In *Robotics: science and systems*. Citeseer, 10.

[5] Itiel Dror and Stevan Harnad. 2008. Offloading cognition onto cognitive technology. John Benjamins Publishing.

[6] Luciano Floridi. 2008. The method of levels of abstraction. *Minds and machines* 18, 3 (2008), 303–329.

[7] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14, 3 (2004), 349–379.

[8] Marlena R Fraune, Selma Šabanović, Eliot R Smith, Yusaku Nishiwaki, and Michio Okada. 2017. Threatening flocks and mindful snowflakes: How group entitativity affects perceptions of robots. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 205–213.

[9] Steve Guglielmo and Bertram F Malle. 2019. Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one* 14, 3 (2019), e0213544.

[10] Alistair M. C. Isaac and Will Bridewell. 2017. White lies on silver tongues: Why robots need to deceive (and how). In *Robot Ethics 2.0*, Patrick Lin, Ryan Jenkins, and Keith Abney (Eds.). 157–172.

[11] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *AAAI Conference on Artificial Intelligence, Ethics, and Society*.

[12] Kheng Lee Koay, Dag Sverre Syrdal, Michael L Walters, and Kerstin Dautenhahn. 2009. A user study on visualization of agent migration between two companion robots. (2009).

[13] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 633–644.

[14] Bradley Oosterveld, Luca Brusatin, and Matthias Scheutz. 2017. Two bots, one brain: Component sharing in cognitive robotic architectures. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 415–415.

[15] Tom Williams, Daniel Ayers, Camille Kaufman, Jon Serrano, and Sayanti Roy. 2021. Deconstructed Trustee Theory: Disentangling Trust in Body and Identity in Multi-Robot Distributed Systems. In *Proceedings of the 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[16] Tom Williams, Daniel Ayers, Camille Kaufman, Jon Serrano, Shania Jo Runningrabbit, Sayanti Roy, Poulomi Pal, Alexandra Bejarano, and Ryan Blake Jackson. 2020. Identity Performance in Multi-Robot Distributed Systems. (2020).

[17] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert Robot-Robot Communication: Human Perceptions and Implications for Human-Robot Interaction. *Journal of Human-Robot Interaction* (2015).

[18] Tom Williams and Matthias Scheutz. 2015. POWER: A domain-independent algorithm for probabilistic, open-world entity resolution. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1230–1235.

[19] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J de Visser. 2020. The Confucian matador: three defenses against the mechanical bull. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 25–33.

[20] K. Winkle, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner. 2021. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In *2021 16th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.