Understanding and Influencing User Mental Models of Robot Identity

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

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

***Abstract*—Research has shown that the relationship between robot mind, body, and identity is flexible and can be performed in a variety of ways. Our research explores how identity performance strategies used among robot groups may be presented through group identity observables (design cues), and how those strategies impact human-robot interactions. Specifically, we ask how group identity observables lead observers to develop different mental models of robot groups, and different perceptions of trust and group dynamics constructs.**

***Index Terms*—Human-Robot Interaction, Robot Groups, Identity Performance Strategies, Mental Models**

I. INTRODUCTION

Typically, Human-Robot Interaction (HRI) research has focused on interactions between humans and *individual* robots. However, HRI domains, such as space exploration [1], [2], [3], search-and-rescue [4], [5], and healthcare [6], [7], involve interactions with *groups* of robots. Interactions with groups of robots are particularly complex due to the number of robot minds (underlying cognitive architectures), bodies (physical constructs), and identities (performed personas) involved. While individual robots typically have humanlike 1-1-1 mind-body-identity associations (i.e. one mind controls one body and presents one identity) from a user's perspective, these associations can break down in robot groups.

Recently, researchers like Williams et al. [8], Luria et al. [9], Reig et al. [10], Tejwani et al. [11], and Jackson et al. [12] have highlighted the flexibility of the relationship between robot mind, body, and identity (i.e. the number of minds, bodies, and identities involved and the associations among them). With this flexibility comes a tension between those relationships at the robot software level versus how they are perceived by humans. For example, Williams et al. [8] consider human interactions with robots that are networked together and controlled by a single mind yet are presented as distinct individuals. This *identity performance strategy* facilitates a user mental model of multiple individuals with distinct minds, bodies, and identities, despite being controlled by a single mind at the software level. Other identity performance strategies (e.g., presentation as a hive mind) may facilitate different mental models (e.g., a single individual whose mind is distributed across multiple bodies).

This distinction between identity performance strategies is critical as the number of bodies and identities involved in a user's mental model dictates where and how they believe trust can be placed, and how they allocate trust to different *trust loci* [8]. This raises the question of *How can robot identity be leveraged to design and enhance human-robot group interactions?* To address this overarching research question, we are first exploring the following underlying questions:

1) How can different identity performance strategies be presented to users?
2) How do different identity performance strategies affect users' social perceptions of robot groups (e.g. trust)?
3) How can different identity performance strategies lead to different mental models of robot groups and their constituent minds, bodies, and identities?

Jackson et al. [12] reason about these questions with respect to robot identity design. This discussion is presented through a *Levels of Abstraction* theoretic account of robot identity in which one must specify the *Level of Abstraction (LoA)* from which a concept is analyzed. An LoA is defined as the set of *observables* available at that LoA. Jackson et al. [12] argue that from a user's perspective (or LoA) the ascription of a unique robot identity may depend on key observables such as naming, speech, and behavior. However, this account of *identity observables* has yet to be validated empirically, and it is unclear what mental models people build about robot groups, or what observables lead to those mental models.

Moreover, it is unclear how these observables and the mental models they evoke impact group dynamics constructs [13] like entitativity [14]. The perceived entitativity of robot groups (i.e., how unified the group appears to be) is a key dimension of group perception that substantively mediates the quality of interaction. For instance, increasing perceptions of robot group entitativity has led to more positive perceptions and willingness to interact with robots [15]. Fraune et al. [14] also emphasizes how entitativity can be manipulated by design choices relating to identity observables.

Thus, overall we aim to understand how robot identity may be leveraged to design and enhance human-robot interactions. Towards this, we explore how identity observables may be used to present identity performance strategies and how those strategies affect human-robot interactions. Specifically, we study the effects of group identity observables on social perceptions like trust, mental model formation, and entitativity.

II. RESEARCH APPROACH

Towards these goals, we conducted two online studies: a replication study (n=189) of Williams et al. [8]'s work and a broad sensemaking study on robot group presentation (n=166).

A. Trust in Robot Bodies and Identities

Williams et al. [8] consider the potential for robot bodies and identities to be perceived as trustworthy to different extents and introduce a new theory of human-robot trust, Deconstructed Trustee Theory. To explore this potential and provide support for their new theory, Williams et al. conducted an online experiment in which participants viewed simulations of two Astrobee robots and rated the perceived trustworthiness of each body and identity. In the videos, different names and voices were used to present different identities and to indicate the migration of an identity between bodies. Participants viewed the robots using one of two Communication Policies (Body-Identity Associating Language or Body-Identity Dissociating Language) and one of two Action Policies (Trust-Building or Trust-Damaging).

A limitation of this study was that the chosen identity observables (names and voices) may have affected the perceived trustworthiness of the robots. For example, the names used, "Honey" and "Bumble", may have led people to assign personality qualities based on the meaning of the names (e.g. Honey being sweet and Bumble being careless). The voices also varied in gender presentation; a critical consideration when designing robot identity, as human gender stereotypes and norms carry over into human-robot communication [16]. We thus conducted a replication study with new (humanoid) names and (gender-controlled) voices.

As in the original experiment, we conducted a Bayesian statistical analysis. While we did not find any significant effects of robot gender and trust, we did find new cases in which each policy and the Locus of Trust (i.e. where trust is placed either body or identity) affected the perceived level of trust in the presented robot bodies and/or identities. The full paper regarding this replication is in preparation for submission to ACM Transactions on HRI.

B. Impact of Robot Group Presentation Strategies on Mental Model Formation

Next, we systematically explored how changes in different *group identity observables* might impact the mental models people develop of robot groups during initial human-robot introductions [17]. We defined and conceptualized five key observables: (1) Speaking (who speaks and when), (2) Self-reference (how an identity speaks of itself or the body it inhabits), (3) Other-reference (how an identity speaks of other identities or the bodies they inhabit), (4) Naming (the name used for oneself, another, or the group), and (5) Name and Voice Distinctiveness (the way body-identity alignment is communicated). These observables may be leveraged at the user's LoA to infer relationships between the minds, bodies, and identities of a robot group and to construct corresponding mental representations. For this study, participants viewed and reflected on animated storyboards across 42 conditions representing key combinations of these five observables and answered open- and close-ended questions.

We used a Grounded Theory-informed analysis [18] of participants' qualitative feedback to identify the types of mental models people developed, and used statistical analysis to show how variations in group identity observables led to those different mental models. Through this analysis, we identified two fundamentally different taxonomies of mental models: Intelligence Distribution and Social Relationships. Additionally, we used participants' quantitative assessments to understand how those varying observables, and the mental models they evoked, influenced perceptions of group entitativity. Specifically, a trend we found across all observables was that shared behavior and qualities resulted in more frequent formation of a One-for-all (i.e. multiple bodies share a single identity/mind) Intelligence Distribution mental model and increased entitativity. While, unique behaviors and qualities resulted in more frequent formation of a One-for-one (i.e. each body has a distinct identity/mind) Intelligence Distribution mental model and decreased entitativity. Overall, we demonstrated the importance of designing robot identity as even the slightest manipulation in the observables lead to changes in the mental models and perceptions observers formed.

III. FUTURE WORK

These studies were our first steps towards understanding how robot identity may be leveraged to design and enhance human-robot group interactions. Both studies explored design techniques that may be utilized to distinctly communicate identity and its association with a robot body and the effect those techniques have on social perceptions, mental model formation, and entitativity. In future work, we aspire to expand on and replicate the results of our preliminary, broad sensemaking study in lab environments, with real robots and a smaller number of experimental conditions with the goal of providing further empirical grounding for the concepts and taxonomies laid out while further addressing our research questions.

We first intend to conduct a series of lab experiments to explore each of the five group identity observables individually. Then, we intend to explore the interactions between identity observables that are most closely tied to each other. This will allow us to understand the effects each individual observable and certain observable pairings have on both mental model formation and on group dynamics constructs. With these experiments, participants can form more concrete and thorough mental models about a robot group as they will be directly interacting with a group of robots. Overall, we hope to provide the HRI community with further understanding of how robot identity can be used to present different identity performance strategies and leveraged to influence user mental models of robot groups.

ACKNOWLEDGMENTS

This work was supported by NASA Early Career Faculty award 80NSSC20K0070.

References

[1] T. Fong, M. Bualat, L. Edwards, L. Flückiger, C. Kunz, S. Lee, E. Park, V. To, H. Utz, N. Ackner *et al.*, "Human-robot site survey and sampling for space exploration," in *Space 2006*, 2006, p. 7425.

[2] J. W. Crandall, M. A. Goodrich, D. R. Olsen, and C. W. Nielsen, "Validating human-robot interaction schemes in multitasking environments," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 4, pp. 438–449, 2005.

[3] M. G. Bualat, T. Smith, E. E. Smith, T. Fong, and D. Wheeler, "Astrobee: A new tool for ISS operations," in *2018 SpaceOps Conference*, 2018, p. 2517.

[4] G.-J. M. Kruijff, F. Colas, T. Svoboda, J. Van Diggelen, P. Balmer, F. Pirri, and R. Worst, "Designing intelligent robots for human-robot teaming in urban search and rescue," in *2012 AAAI Spring Symposium Series*, 2012.

[5] I. R. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human-robot teaming for search and rescue," *IEEE Pervasive Computing*, vol. 4, no. 1, pp. 72–79, 2005.

[6] A. Di Nuovo, F. Broz, N. Wang, T. Belpaeme, A. Cangelosi, R. Jones, R. Esposito, F. Cavallo, and P. Dario, "The multi-modal interface of robot-era multi-robot services tailored for the elderly," *Intelligent Service Robotics*, vol. 11, no. 1, pp. 109–126, 2018.

[7] S. H. Alsamhi and B. Lee, "Blockchain for multi-robot collaboration to combat covid-19 and future pandemics," *arXiv preprint arXiv:2010.02137*, 2020.

[8] T. Williams, D. Ayers, C. Kaufman, J. Serrano, and S. Roy, "Deconstructed trustee theory: Disentangling trust in body and identity in multi-robot distributed systems," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 262–271.

[9] M. Luria, S. Reig, X. Z. Tan, A. Steinfeld, J. Forlizzi, and J. Zimmerman, "Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, 2019, pp. 633–644.

[10] S. Reig, E. J. Carter, T. Fong, J. Forlizzi, and A. Steinfeld, "Flailing, hailing, prevailing: Perceptions of multi-robot failure recovery strategies," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 158–167.

[11] R. Tejwani, F. Moreno, S. Jeong, H. W. Park, and C. Breazeal, "Migratable ai: Effect of identity and information migration on users' perception of conversational ai agents," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 877–884.

[12] R. B. Jackson, A. Bejarano, K. Winkle, and T. Williams, "Design, performance, and perception of robot identity," in *Workshop on Robo-Identity: Artificial identity and multi-embodiment at HRI 2021*, 2021.

[13] A. M. Abrams and A. M. Rosenthal-von der Pütten, "I–c–e framework: Concepts for group dynamics research in human-robot interaction," *International Journal of Social Robotics*, pp. 1–17, 2020.

[14] M. R. Fraune, S. Šabanović, E. R. Smith, Y. Nishiwaki, and M. Okada, "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, 2017, pp. 205–213.

[15] M. R. Fraune, B. C. Oisted, C. E. Sembrowski, K. A. Gates, M. M. Krupp, and S. Šabanović, "Effects of robot-human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact," *Computers in Human Behavior*, vol. 105, p. 106220, 2020.

[16] R. B. Jackson, T. Williams, and N. Smith, "Exploring the role of gender in perceptions of robotic noncompliance," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 559–567.

[17] A. Bejarano, S. Reig, P. Senapati, and T. Williams, "You had me at hello: The impact of robot group presentation strategies on mental model formation," in *Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022.

[18] K. Charmaz, *Constructing grounded theory*. sage, 2014.