Identity Performance in Multi-Robot Distributed Systems

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

Daniel Ayers
MIRRORLab
Colorado School of Mines
Golden, CO USA
dayers@mines.edu

Camille Kaufman
MIRRORLab
Colorado School of Mines
Golden, CO USA
cjkaufman@mines.edu

Jon Serrano
MIRRORLab
Colorado School of Mines
Golden, CO USA
serrano@mines.edu

Shania Jo Runningrabbit
MIRRORLab
Colorado School of Mines
Golden, CO USA
srunningrabbit@mines.edu

Sayanti Roy
MIRRORLab
Colorado School of Mines
Golden, CO USA
sayantiroy@mines.edu

Poulomi Pal
MIRRORLab
Colorado School of Mines
Golden, CO USA
poulomipal@mines.edu

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

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

***Abstract*—In this work we explore the relationship between mind, body, and identity in multi-robot distributed systems. Specifically, we explore how robot designers can adapt robots to selectively perform identities, and the effects this may have on human-robot trust, especially with respect to the novel concepts of trust localization, dissociation, and fragmentation.**

***Index Terms*—robot identity performance, multi-robot systems, distributed systems, human-robot trust**

I. Introduction

The Lunar Orbital Platform-Gateway will serve as a staging point for crewed and uncrewed missions to the Moon, Mars, and beyond [5]. While the Gateway will sustain human crews for small periods of time, it will be primarily staffed by autonomous caretaker robots like the free-flying Astrobee platform [30]: the Gateway's sole residents during quiescent (uncrewed) periods [4]. This creates a unique human-technical system comprised of two categories of human teammates: ground control workers permanently stationed on earth and astronauts that may transition over time between work on Earth, the Gateway, the Moon, and Mars; and three types of machine teammates: robot workers stationed on the Gateway; robot workers stationed on the Moon and Mars; and the Gateway itself. In this paper we analyze the unique nature of *robot identity* in this type of multi-robot system, the *performative* nature of identity in such systems, and the unique opportunities and challenges it presents, especially with respect to human-robot trust. To do so, we must first examine the relation between mind, body, and identity.

II. Mind, Body, and Identity

Thought experiments like Dennett's "Where Am I" [6] (see also [7], [14]) have led to prolific discussion in the Philosophy of Mind and Metaphysics literatures on the relation between Mind, Body, and Self [21], [26], [29], and speculation as to whether some far-future technology will allow these three concepts to be dissociated in humans, (i.e., remove these concepts' currently necessary 1-1 mapping), and how cognitive technologies may already be distributing human cognition [9].

Meanwhile, robotic mind, body, and self are already dissociated in deployed robotic systems, with serious implications for human-robot teaming. While modern robots are presented as monolithic systems with one mind, body, and identity, this is rarely the case in practice. NASA's Astrobees have discrete bodies and names, but their "mind", i.e., the computation governing their behavior, is distributed across multiple machines. In fact, the Gateway and its Astrobees can be viewed as a single integrated system with a single "mind" but multiple bodies, each with a unique human-assigned identity.

Moreover, HRI researchers are increasingly blurring the distinction between mind, body, and identity, through architectural mechanisms like component sharing. Oosterveld et al. [24], for example, present a pair of robots with separate perception and motion systems but shared dialogue and goal management components. This enables each robot to report what the other robot sees, and pass along information and commands to the other robot, while maintaining (or as we will argue, performing) a unique identity.

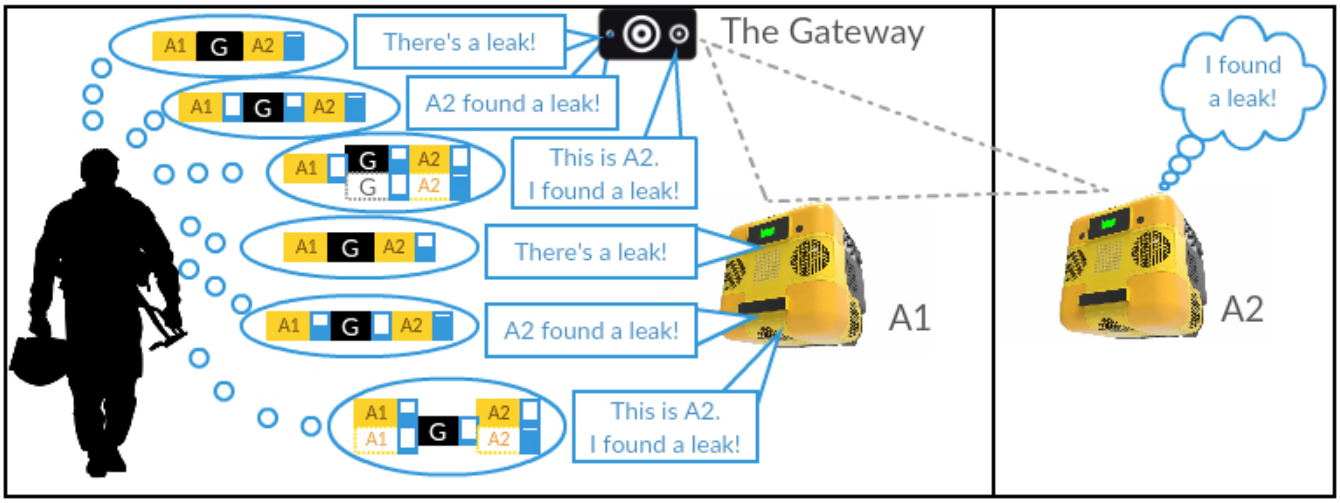

Fig. 1. Identity Performance Strategies, showing how different levels of human-robot trust might be built up in different robot bodies and identities, both holistically and individually.

Identity is strongly associated with a body through naming [8]. We argue that naming *reifies* identity both philosophically and cognitively. Because naming presupposes a named referent, it should trigger cognitive processses in humans such as reference resolution and hypothesization [36], [37], thus creating and concretizing mental representations associated with the robot's identity. We argue that these mental representations of identity can then be reinforced by various robot behaviors. Robots that refer to each other by name, for example, may draw a distinction between themselves and others, reinforcing the notion of distinct robot identities associated with different robot bodies.

Robots that communicate may reinforce whatever individual or collective identity they have previously presented, as communication will trigger listeners to try to identify *who* is speaking. Similarly, robots that perform blameworthy actions may reinforce whatever individual or collective identity they have previously presented, as performance of those actions will trigger observers to try to identify *who* they should blame. And robots that describe individual goals and beliefs may reinforce representations of individual identities.

III. Agency and Identity

This analysis highlights that for robot identity what is truly important is how identity is *perceived* by users. This is in many ways similar to the role of agency in Human-Robot Interaction. As defined by Floridi and Sanders, a thing is an agent if it is interactive, autonomous, and adaptable; properties that can be analyzed at different levels of abstraction, defined by what is observable from different perspectives [11]. From a developer's perspective, a robot may not be an agent, because the developer can observe the algorithms behind a robot's behavior, and determine that, for example, a robot may not truly be adaptable. This insight may be based on observations that, for example, a robot's changes in behavior are purely changes of state rather than learning-based updates to the transition functions that determine those changes of state.

From a user's perspective, however, that same robot can be an agent, because without knowledge of the robots underlying algorithms, the robot can satisfy those required properties from the user's perspective. Thus, what is most significant for human-robot interactions is not whether a robot is an agent from the developer's perspective, but rather whether a robot is an agent from users' perspectives, because that perception is what will actually impact interactions. It may be the user's perception of a robot's agency at their own level of abstraction, for example, that gives that robot persuasive power, and not the decision by developers or philosophers as to whether the robot has agency at their levels of abstraction.

This Levels-of-Abstraction account of robot agency then allows us to make new insights about robot identity as well. Whereas agency is concerned with whether an individual is capable of action, identity is concerned with whether there is an individual to be perceived and modeled to begin with. We argue that identity can also be viewed from different levels of abstraction. From a developer's perspective, a robot in a multi-robot distributed system may not be a unique individual, as any interactions ostensibly engaged in with that robot will in fact be interactions between the user and the entire multi-robot distributed system, even if this is non-obvious or unobserveable from the user's Level of Abstraction. However, from the user's perspective, the robot of course may well be perceived to be a unique individual, and, as with agency, we would expect that it is this perception that will impact interaction moreso than the robot's status at the developer's Level of Abstraction.

IV. Identity Performance

We argue that for robots, identity-body association, and the extent to which this association is observable at the user's Level of Abstraction, are design choices to be made by robot designers (or by robots themselves). We further argue that the fact that identity-body association may differ at the developer and user's Levels of Abstraction based on robots' designed or selected behaviors makes robot identity *performative*. That is, robots with unique names and identities may best be viewed as performing identities for human benefit; a performance that may be dropped, or whose illusion may be broken, at any time.

Body-identity alignment has not been previously understood as a communicative design choice: while within the HRI community there has been significant research on agent migration, which focuses on the ability of agents to "hop" between bodies [13], [15], [22], [23], with an emphasis on the ability for those agents to maintain and project consistent and coherent identities [17], [18], [25], that work focuses on permanent migration of distinct agents (e.g., when one robot breaks down) rather than the design choices that will affect how identity is selectively performed in the communication of individual utterances. Accordingly, almost nothing is known about the implications of this design choice, beyond the potential benefits we have suggested above.

This performativity is a key design tool for robot designers. In fact, a number of performative design patterns have been proposed over the past several years. One example is Kwon et al.'s work on expressing robot incapability [19], in which a robot pretends to physically struggle to communicate that an object is heavy, using the humanlike metaphor of muscle strain to effectively communicate using a robot morphology that cannot actually experience this sort of strain. Another example is Williams et al.'s work on performative communication in multi-robot systems [34], in which robots verbally communicate human-relevant information between themselves, in order to keep humans apprised of their conversation and keep said humans at ease, even though this is of course not the primary channel through which the robots are actually exchanging information.

We argue that the performance of identity may be particularly useful to designers due to the benefits that may be enabled by the perception of individual, nameable identities. Naming (1) enables humans to more easily refer to robots through natural language (and to do so without specifying identity conditions every time they refer [28]); (2) conveys a sense of value and is viewed as "deserved" even for robots with limited social agency [32]; and (3) increases perceived agency [2] and human-likeness [32], which accordingly increases perceived social bonds [27], mediates decision making [31], and as we argue in our own work, gives robots unique persuasive power [16]. Performing unique robot identities may prevent uncanny valley effects [33], [35]. And, we argue, performing unique robot identities may create new *trust loci*, enabling the development of trust to build up in those unique identities, as the perception of a unique robot identity may lead users to attempt to engage in social cognitive reasoning about those identities, e.g., with respect to whether those identities should be trusted.

The existence of multiple loci of trust (e.g., a robot's body versus the identity performatively associated with it) suggests that while trust in these associated loci will likely be correlated, different levels of trust may ultimately be gained and lost in each of these loci. Accordingly, we argue that, when measuring, modeling and manipulating human-robot trust, it is necessary to deconstruct the robot trustee using a representation composed of discrete loci of trust, including the trustee's body and identity (or bodies and identities, in cases of performative re-embodiment or co-embodiment), and that this is especially important in multi-robot systems.

Furthermore, we argue that performance of identity is a default design choice that could be intentionally subverted through robot communication policies that "break the illusion" of 1-1 body-identity association (e.g., changing the identity performed by a particular robot body for the sake of convenience), and that different robot identity performance strategies might lead to different levels of trust being built in different trust loci, or in the evocation or suppression of different potential loci.

Finally, we argue that different types of trust-affecting actions may have different effects on trust built in different loci. Recent moral psychological work from Guglielmo et al. suggests that human blame is more intense and more subtly differentiated than human praise [12]. Based on this recent evidence, we might similarly expect that impacts on trust in response to trust damaging actions might be more intense than impacts on trust in response to trust building actions. We might also expect that users might be more deliberative and selective about who (i.e., which locus of trust) they are losing trust in for trust-damaging (blameworthy) actions than for trust-building (praiseworthy) actions. These expectations would further suggest that trust-damaging actions might lead to stronger evocation of loci of trust, and stronger drops in the trust built in those loci.

V. Identity Performance on the Gateway

Let us now consider how the identity performance design strategies described above might play out aboard the Gateway. Because the Gateway and its robotic workers will be integrated into a single system, when humans interact with different robot bodies aboard the Gateway, they will in fact be interacting with a single integrated system. Accordingly, we argue that the distinct identities presented by the Gateway and its robots are in fact performed for human benefit. Accordingly, when the integrated Gateway system needs to communicate with a human teammate, it may choose what body to use and what identity to perform. Fig. 1 shows an example scenario in which the integrated Gateway system must make an identity performance choice. Here, Astrobee 2 detects a leak, and wishes to communicate this to a human astronaut co-located with Astrobee 1.

This can be achieved in at least six ways, each of which may differently shape the astronaut's trust in the integrated Gateway system. First, a body must be chosen through which to communicate: the Gateway itself, or the Astrobee co-located with the astronaut. Next, for each of these choices of body, there are three different choices of identity. (1) The chosen body may simply state the information to be communicated (e.g., "There's a leak in the logistics module."). This may facilitate a holistic model of trust where trust is placed in the integrated system as a whole. (2) The chosen body may state the information, and where it comes from (e.g., "Astrobee 2 says there's a leak in the logistics module."). This may facilitate a model of trust where trust is separately allocated to each body-identity pair. (3) The chosen body may performatively communicate from the perspective of the remote robot (e.g., "This is Astrobee 2. There's a leak in the logistics module."). This may facilitate a model of trust where trust is separately allocated to each body and to each identity.

What is more, a variety of further design choices may be opened up as the size of human-robot teams increases. We might expect that as the size of a multi-robot team increases, the cognitive cost of needing to remember the names of and develop relationships with each robot body will become increasingly untenable, especially if robots have identical morphologies and capabilities. As such, designers may choose to explore a variety of group identity performance strategies that go beyond the fiction of n interactive robots with n identities, such as n interactive robots with one identity (a hive-mind), n interactive robots with n copies of one identity (as with Amazon's Alexa), or n robots comprised of m interactive robots with m identities and $n-m$ non-interactive robots without identities (e.g., an earthbound robot that serves as an interface for ground control workers to non-interactive identityless robots on the surface of the Moon).

To understand how these different design choices might impact human-robot trust, we define three novel concepts: trust localization (where is trust placed?), trust dissociation (are body-trust and identity-trust correlated?) and trust fragmentation (are body/identity trust and holistic trust correlated?). Because so little is known about the nature of trust in distributed, integrated, autonomous systems, it is not yet clear how different identity performance strategies will impact these aspects of trust distribution. Moreover, it is not yet clear how these aspects will impact team performance, e.g., whether trust fragmentation and dissociation help or harm long-term performance when team composition changes; or what uncanny valley effects might arise from communication strategies that dissociate body and identity.

VI. Current and Future Directions

In recent work, we have begun to investigate the impact of different identity performance strategies on trust localization and dissociation, through experiments conducted using NASA's Astrobee simulator, using simulated versions of the Bumble and Honey Astrobee robots (see Fig. 2), in the context of *inspection tasks*. Bualat et al.'s report on the

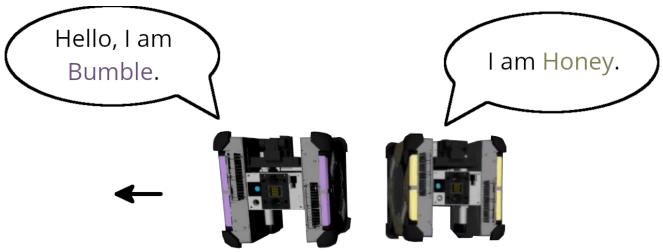

Fig. 2. Introductory dialogue used by simulated robots in recent experimental work

Astrobee system highlights four key inspection tasks that the Astrobee will need to fulfill on the Gateway [4]: spot-checks conducted on-demand at human request; surveys conducted on a periodic schedule; change detection to identify developing problems based on survey and spot-check results; and problem localization to pinpoint the location of anomalous readings and detected changes. All four inspection tasks can be performed with respect to multiple types of critical sensor readings, including noise [1], radiation [3], CO_2 [20], and object positions (detected as RFID signal strength through Astrobee's REALM-2 payload [10]).

While a full account of that work is beyond the scope of this narrowly defined workshop paper, it is nonetheless informative to briefly summarize the results of that work here. First, our results suggest that identity performance strategies can have large effects on trust localization (as can the type of action communicated about by robots (i.e., trust-building vs. trust-damaging actions)), especially where humans believe to be appropriate loci for capability-based trust. Second, our results suggest that while human-robot trust can indeed measurably dissociate between different trust loci, this dissociation is not triggered by different identity performance strategies, but rather by the type of action communicated about by robots (i.e., trust-building vs. trust-damaging actions) and the role played by robots (i.e., actor vs communicator), especially for reliability-based trust.

In future work, we plan to further interrogate the new theories and concepts presented in this paper, including (1) the design space of identity performance strategies and the impacts of those strategies, (2) the factors that impact trust localization, dissociation, and fragmentation, (3) and the resulting effects of trust localization, dissociation, and fragmentation.

Acknowledgments

This work was funded in part by an Early Career Faculty Award from NASA and in part by a Young Investigator Award from the Air Force Office of Scientific Research.

References

[1] James E Allen, Curry I Guinn, and Eric Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.

[2] Thomas Arnold and Matthias Scheutz. Hri ethics and type-token ambiguity: what kind of robotic identity is most responsible? *Ethics and Information Technology*, pages 1–10, 2018.

[3] Thomas Berger, Bartos Przybyla, Daniel Matthiä, Günther Reitz, Sönke Burmeister, Johannes Labrenz, Pawel Bilski, Tomasz Horwacik, Anna Twardak, Michael Hajek, et al. DOSIS & DOSIS 3D: long-term dose monitoring onboard the columbus laboratory of the international space station (ISS). *Journal of Space Weather and Space Climate*, 6, 2016.

[4] Maria G Bualat, Trey Smith, Ernest E Smith, Terrence Fong, and DW Wheeler. Astrobee: A new tool for iss operations. In *2018 SpaceOps Conference*, 2018.

[5] Jason C Crusan, R Marshall Smith, Douglas A Craig, Jose M Caram, John Guidi, Michele Gates, Jonathan M Krezel, and Nicole B Herrmann. Deep space gateway concept: Extending human presence into cislunar space. In *2018 IEEE Aerospace Conference*, pages 1–10. IEEE, 2018.

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

[7] Daniel C Dennett. *Brainstorms: Philosophical essays on mind and psychology*. MIT press, 1981.

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

[9] Itiel Dror and Stevan Harnad. Offloading cognition onto cognitive technology. In *Cognition Distributed: How Cognitive Technology Extends Our Minds*. John Benjamins Publishing, 2008.

[10] Patrick W Fink, Timothy F Kennedy, Lazaro Rodriguez, James L Broyan, Phong H Ngo, Andrew Chu, Ami Yang, Donald M Schmalholz, Robert W Stonestreet, Robert C Adams, et al. Autonomous logistics management systems for exploration missions. In *AIAA SPACE and Astronautics Forum and Exposition*, 2017.

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

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

[13] Wan Ching Ho, Kerstin Dautenhahn, Mei Yii Lim, Patricia A Vargas, Ruth Aylett, and Sibylle Enz. An initial memory model for virtual and robot companions supporting migration and long-term interaction. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 277–284. IEEE, 2009.

[14] Douglas R Hofstadter and Daniel C Dennett. *The Mind's I: Fantasies and Reflections on Self & Soul*. Basic Books, 2006.

[15] Michita Imai, Tetsuo Ono, and Tameyuki Etani. Agent migration: communications between a human and robot. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, volume 4, pages 1044–1048. IEEE, 1999.

[16] Ryan Blake Jackson and Tom Williams. On perceived social and moral agency in natural language capable robots. In *Proceedings of the 2019 HRI Workshop on The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI*, 2019.

[17] Kheng Lee Koay, Dag Sverre Syrdal, Michael L Walters, and Kerstin Dautenhahn. A user study on visualization of agent migration between two companion robots. In *Proceedings of the 13th Conference on Human-Computer Interaction*, 2009.

[18] Michael Kriegel, Ruth Aylett, Kheng Lee Koay, KD Casse, Kerstin Dautenhahn, Pedro Cuba, and Krzysztof Arent. Digital body hopping-migrating artificial companions. *Digital Futures*, 2010.

[19] Minae Kwon, Sandy H Huang, and Anca D Dragan. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95, 2018.

[20] Thomas F Limero and William T Wallace. What air and water quality monitoring is needed to protect crew health on spacecraft? *New Space*, 5(2):67–78, 2017.

[21] William G Lycan. Consciousness as internal monitoring, I: the third philosophical perspectives lecture. *Philosophical Perspectives*, 9:1–14, 1995.

[22] Pauli Misikangas and Kimmo Raatikainen. Agent migration between incompatible agent platforms. In *Proceedings 20th IEEE International Conference on Distributed Computing Systems*, pages 4–10. IEEE, 2000.

[23] Tetsuo Ono, Michita Imai, and Ryohei Nakatsu. Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism. *Advanced Robotics*, 14(4):311–326, 2000.

[24] Bradley Oosterveld, Luca Brusatin, and Matthias Scheutz. Two bots, one brain: Component sharing in cognitive robotic architectures. In *Companion Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction*, pages 415–415. ACM, 2017.

[25] Samantha Reig, Jodi Forlizzi, and Aaron Steinfeld. Leveraging robot embodiment to facilitate trust and smoothness. In *Companion Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 742–744. IEEE, 2019.

[26] Giuseppe Riva, F Davide, and WA IJsselsteijn. Being there: The experience of presence in mediated environments. *Being there: Concepts, effects and measurement of user presence in synthetic environments*, 5, 2003.

[27] Matthias Scheutz. The inherent dangers of unidirectional emotional bonds between humans and social robots. In *Robot ethics: The ethical and social implications of robotics*, page 205. MIT Press, 2011.

[28] John R Searle. Proper names. *Mind*, 67(266):166–173, 1958.

[29] Lawrence Shapiro. *Embodied cognition*. Routledge, 2010.

[30] Trey Smith, Jonathan Barlow, Maria Bualat, Terrence Fong, Christopher Provencher, Hugo Sanchez, and Ernest Smith. Astrobee: A new platform for free-flying robotics on the international space station. In *Proceedings of International Symposium on Artificial Intelligence, Robotics and Automation in Space*, 2016.

[31] Megan Strait, Gordon Briggs, and Matthias Scheutz. Some correlates of agency ascription and emotional value and their effects on decision-making. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 505–510. IEEE, 2013.

[32] Ja-Young Sung, Lan Guo, Rebecca E Grinter, and Henrik I Christensen. "My Roomba is Rambo": intimate home appliances. In *International Conference on Ubiquitous Computing*, pages 145–162. Springer, 2007.

[33] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. From one to another: How robot-robot interaction affects users' perceptions following a transition between robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 114–122. IEEE, 2019.

[34] Tom Williams, Priscilla Briggs, and Matthias Scheutz. Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction*, 4(2):24–49, 2015.

[35] Tom Williams, Priscilla Briggs, and Matthias Scheutz. Covert robot-robot communication: Human perceptions and implications for human-robot interaction. *Journal of Human-Robot Interaction*, 2015.

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

[37] Tom Williams and Matthias Scheutz. Reference in robotics: A givenness hierarchy theoretic approach. *The Oxford handbook of reference*, 2019.