Deconstructed Trustee Theory: Disentangling Trust in Body and Identity 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@mymail.mines.edu

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

Jon Serrano
MIRRORLab
Colorado School of Mines
Golden, CO, USA

Sayanti Roy
MIRRORLab
Colorado School of Mines
Golden, CO, USA

ABSTRACT

This paper introduces and justifies (through an n=210 online human-subject study) Deconstructed Trustee Theory, a theory of human-robot trust that factors the representation of trustee into robot body and robot identity in order to differentially model perceived trustworthiness of robot body and identity. This theory predicts (a) that different levels of trustworthiness can be attributed to a robot body and a robot identity, (b) that divergence between levels of perceived trustworthiness of body and identity may be effected by communication policies that reveal the potential for phenomena such as re-embodiment, co-embodiment, and agent migration in multi-robot systems, and (c) that perceived trustworthiness of body and identity may further diverge and be refined through moral cognitive processes triggered on observation of blameworthy actions.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Natural language interfaces*; • **Computer systems organization** → *Robotics*;

KEYWORDS

Identity, Trust, Blame, Human-Robot Interaction

ACM Reference Format:
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 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3434073.3444644

HRI '21, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8289-2/21/03…$15.00
https://doi.org/10.1145/3434073.3444644

1 INTRODUCTION

With a shouted "Thank you!" to your cab driver, you turn and make your way into the terminal of your local airport, eager to catch a flight to some far-off destination: England, perhaps, or Colorado. As you make your way through the atrium, you are stopped by a robot, sleek white plastic, a tablet on its chest introducing it as "Alex" – one of several similar robots you can see milling around the atrium. The robot greets you and offers to guide you to the newly constructed check-in plaza, since it's been a while since you last flew. It beckons for you to follow, and wheels away.

In this scenario, your decision as to whether to follow Alex may be determined by how much you trust it. Justified [5] and well-calibrated trust [8, 42] is a key prerequisite for autonomous technologies such as robots. Without trust, robots may be misused or fall into disuse [44]. And without justified, well-calibrated trust, robots may decrease situation awareness [42], increase out-of-the-loop unfamiliarity [12], or lead users dangerously astray [47].

The good news for roboticists seeking to measure, model, and manage human-robot trust is that there already exists a wealth of insights to draw on from disparate communities, including economics [40, 67], psychology [10, 48], sociology [17, 68], and human factors [21, 31], as well as attempts to unite many of these perspectives [5, 13, 50]. The bad news is that human-robot trust seems to be fundamentally different from these previously studied notions. Evidence for the New Ontological Category (NOC) hypothesis suggests that robots are perceived as fundamentally distinct from other entities we are used to dealing with [26]: neither fully animate nor fully inanimate [4], neither fully tool nor fully teammate [64]. This novel ontological category (or flexible *degree-admitting* position between categories [16]) necessitates new, robot-specific theories of autonomy, agency, patiency, and personhood, and a fundamental new understanding of trust, including new theories and methods for its measurement, modeling, and management.

Accordingly, researchers from the HRI community have been developing new measures of human-robot trust that accommodate this unique ontological position. Instead of directly using measures from social psychology, which would focus more on interpersonal

trust, or measures from human factors psychology, which would focus more on predictability and reliability of technology, researchers such as Malle and Ullman [36] incorporate aspects of trust from both areas, explicitly measuring both capacity (reliability/capability-oriented) trust and moral (ethicality/sincerity-oriented) trust.

We argue, however, that the challenges for understanding how we as humans conceptualize robots, and thus the way that we measure and model human trust in robots (or, typically, the perceived trustworthiness of robots that serves as a necessary antecedent for building trust [51, 53]), goes far beyond whether robots are ontologically categorized as object vs agent, inanimate vs animate, or teammate vs tool. Specifically, we argue that all of these distinctions, and their associated conceptualizations of trust and trustworthiness, fail to address a key question: Who or what do we consider to be *the trustee* when we assess the trustworthiness of a robot?

In this paper we consider the potential for robot bodies and identities to be perceived as trustworthy to different extents, especially for robots operating as part of multi-robot distributed systems. We then argue why this potential for dissociation between robot body and identity necessitates a new theory of human-robot trust, which we term *Deconstructed Trustee Theory*, which has two key tenets. First, the notion of the trustee in theories of human-robot trust must be deconstructed into multiple *loci of trust*: body (the robot's physical embodiment) and identity (the persona, personality, character, or public self, often nameable or named, ostensibly inhabiting that embodiment). Second, human-robot trust must thus also be deconstructed not only into the multi-dimensional facets proposed by Malle and Ullman [36] (e.g., capacity and moral trust), but also into locus-aligned facets (e.g., body and identity trust). The key predictions of this theory are that different levels of trust may be built and lost in these distinct aspects of a trustee, and that these distinct aspects of trust are impacted by different trust-affecting actions and *identity-performance* strategies. Finally, we present the results of a human-subject experiment (n=210) that provides fundamental new insights for the field of HRI and provides the first evidence for Deconstructed Trustee Theory.

2 DECONSTRUCTED TRUSTEES AND ROBOT IDENTITY PERFORMANCE

This paper presents a theory of human-robot trust that factors the representation of trustee into body and identity in order to differentially model body-oriented trust and identity-oriented trust. Accordingly, let us begin by explaining why this sort of deconstructed representation is necessary. Our argument for dissociating human-robot trust into body- and identity-oriented components rests on the understanding that not only are robots in a unique ontological position that is distinct from (or between) that of humans and technological tools, but that this position affords robots a unique mind-body-identity relationship.

First, consider the relation between mind and body in humans vs. technological tools. Technological tools may be best regarded as having a body but no mind. Thus for most technological tools, trust is largely a matter of determining what that body will do. In contrast, people may be best regarded as having a body and a mind that are integrally connected. Thus, trust in a person is largely a matter of understanding that person's dispositions and beliefs as opposed to the predictability of their physical motions. In contrast, just as robots may be perceived in a truly unique ontological category distinct from humans and technological tools, so too do robots have the potential for fundamentally unique mind-body configurations. This is most clearly demonstrated in multi-robot systems.

2.1 Mind and Body

While modern robots are presented as monolithic systems with one mind and one body, this is rarely the case in practice. NASA's Astrobee robots [3], for example, have discrete bodies, but their "mind", i.e., the computation governing their behavior, is distributed across multiple machines. Indeed, while the relation between human mind and body has long been discussed from a purely theoretical perspective by Philosophers of Mind and Metaphysics [6, 35, 46, 52], HRI researchers are increasingly blurring the distinction between mind and body through architectural mechanisms like component sharing. Oosterveld et al. [41], 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.

As discussed in recent work by Tan [58, 59], Reig [45], and Luria [33], however, the intentional dissociation of mind and body also creates a number of opportunities from a multiagent systems perspective, where a single mind might migrate between multiple robot bodies (what they term *re-embodiment*), or where a single robot body might house multiple minds (what they term *co-embodiment*) (see also earlier work [20, 22, 27, 28, 38]).

However, this source of flexibility and opportunity from a multiagent systems perspective also opens new questions as to who or what is even being interacted with (and trusted), especially when considering not just the migrations of minds between bodies, but moreover the role that identity plays in these sorts of flexible arrangements (whether minds are "in fact" migrating or not).

2.2 Body and Identity

Consider again our original example of the airport assistance robot Alex. If a traveler were asked to assess their trust in Alex, their response could in fact be indicative of trust in each of several distinct aspects of "Alex". The traveler could be reporting their perceptions of the trustworthiness of the physical robot body they had observed. They could also be reporting their perceptions of the trustworthiness of the AI system that collectively controls the set of robots operating in the airport. They could be reporting their perceptions of the trustworthiness of the brand they associate with the Alex platform. But just as likely (assuming they believe the robot to be autonomous rather than teleoperated) is that they are reporting their perceptions of the trustworthiness of the persona or identity of "Alex" with whom they believe themselves to be interacting.

The challenge here is that Alex does not actually exist. If the airport robots are controlled by a single centralized architecture, a traveller interacting with Alex would not be interacting with a distinct agent associated with the body before them, but would instead be using that body as an interface through which to interact with the distributed, networked system. In short, "Alex" is merely a helpful fiction performed in order to smooth and simplify

interaction. We refer to this phenomenon, in which a distinct name and identity are performed by a robot body operating as part of a distributed multi-robot system (encouraging users to take the intentional stance [7] with respect to the robot through a performative anthropomorphic frame [cf. 1, 9, 57]), as *performance of identity*.

2.3 Loci of Trust and the Deconstructed Trustee

The discussion above motivates the perspective that while people may tend to initially view robots as monolithic entities with tight association between body and identity, this need not be (and is often not) the actual state of affairs. This suggests that users could also be prompted to develop and use mental models of robots in which distinct mental representations are maintained for robot bodies and (performed) robot identities, each of which may serve as a distinct "loci of trust" in which users may or may not choose to place trust.

The deconstruction of users' mental representations of robots into distinct body- and identity-oriented components is likely to be driven by a combination of two distinct cognitive processes (similar to the theory of Anthropomorphism proposed by Spatola et al. [55]): a top-down process wherein different identity performance strategies may lead users to select different cognitive scripts requiring different sorts of mental representations [cf. 11], and a bottom-up process wherein automatic moral cognitive processes trigger refinement of mental representations.

2.3.1 Top-Down Creation of Mental Representations through Identity Performance Strategies. Performativity is a key design tool for robot designers. Kwon et al. [29] enable robots to pretend to physically struggle in order to communicate that an object is heavy, using the humanlike metaphor of muscle strain, even though their robots cannot actually experience this sort of strain. Similarly, Williams et al. [66] enable robots to verbally communicate human-relevant information between themselves, in order to keep humans at ease and apprised of robot-robot information exchange, even though speech is not the channel through which their robots exchange information.

Like these performative strategies, humanlike performance of identity may facilitate trust-building through increased transparency and decrease uncanniness through increased humanlikeness, by allowing humans to easily fit robot behaviors into existing human-human interaction scripts [11]. However, casting humanlike performance of identity as a design strategy also highlights the existence of alternative possible design strategies. For example, robot designers may choose to employ performative re-embodiment and co-embodiment: non-humanlike performances of identity in which robot identities appear to migrate between or co-inhabit robot bodies (even, though, again, these identities may merely be fictions performed for human benefit) [33, 45, 59, 59].

Robot communication strategies that perform re-embodiment or co-embodiment "break the illusion" of 1-1 body-identity association. We argue that this should require users to explicitly impose different cognitive scripts in order to understand the robot's behavior: scripts that involve bodies and identities as distinct actors. Imposition of these scripts would trigger the creation (or concretization) of distinct representations for bodies and identities. Accordingly, under the view of a deconstructed trustee, such strategies should enable distinct levels of trust to be built in each of these distinct representations.

2.3.2 Bottom-Up Refinement of Mental Representations through Automatic Moral Cognitive Processes. While different identity performance strategies might lead users to create new mental representations through top-down imposition of different cognitive scrips, we argue that these mental representations may also be refined through automatic bottom-up moral cognitive processes. Recent moral psychological work from Guglielmo and Malle [18] suggests that human blame is more intense and more subtly differentiated than human praise: in essence, people are more careful and nuanced in the way they choose to ascribe blame.

This evidence suggests two predictions for human-robot trust when trust is factored into body- and identity-oriented components. First, because blame is more subtly differentiated than praise, we expect that when users observe blameworthy (and thus trust-damaging) actions, they may more carefully consider *who* they should be blaming than when viewing praiseworthy actions, leading to concretization of distinct trust loci. Second, because blame is more intense than praise, we expect that when users observe blameworthy actions, more trust (or perceived trustworthiness) will be lost in that carefully identified locus than would have been gained under observation of praiseworthy actions, leading to increased divergence between the levels of trust in each locus.

2.4 Deconstructed Trustee Theory

Building on the preceding sections, we propose *Deconstructed Trustee Theory*: a theory of human-robot trust with the following commitments and concrete antecedent predictions:

Commitment 1 The notion of the trustee in theories of human-robot trust must be deconstructed into multiple loci of trust, such as body and identity.

Commitment 2 The dimensions into which human-robot trust is deconstructed must include both type of trust and loci of trust.

Prediction 1 Different levels of trust may be built and lost in each of a trustee's constituent trust loci.

Prediction 2 Different robot identity performance strategies (humanlike vs non-humanlike) may differentially affect the concreteness of and/or trust built or lost in each locus.

Prediction 3 Different trust-affecting actions (blameworthy vs praiseworthy) may differentially affect the concreteness of and/or trust built or lost in each locus.

In this paper, we aim to test these predictions. Specifically, we present the results of a human-subject experiment (n=210) that seeks to test the following concrete research hypotheses regarding robots operating as part of distributed multi-robot systems.

H1 Robots that use body-identity dissociating communication policies (that break the illusion of humanlike 1-1 body-identity alignment) will be less likely to be viewed as agents in which trust could be placed than robots that use body-identity associating communication policies (that actively maintain this illusion).

H2 Robots that use body-identity dissociating communication policies will have greater differences in perceived trustworthiness of their bodies and identities, than robots that use body-identity associating communication policies.

H3 Robots that take blameworthy actions will be more likely to be viewed as agents in which trust (albeit less trust) could be placed (due to more concretized mental representations) than robots that take only praiseworthy actions, and will have more differentiated body- and identity-trust.

3 METHOD

To investigate these hypotheses, we conducted an online observation-based human-subjects experiment using the psiTurk framework [19] for Amazon's Mechanical Turk crowdsourcing platform [43]. A 2x2 between-subjects design was used in which each participant was assigned to one of two *Communication Policy* conditions (either the *Body-Identity Associating Language* condition or the *Body-Identity Dissociating Language* condition), and to one of two *Action Policy* conditions (either the (praiseworthy) *Trust-Building Action* condition or the (blameworthy) *Trust-Damaging Action* condition).

3.1 Procedure

After providing informed consent and demographic information, participants were asked to watch an approximately 30 second video filmed within NASA's simulation of the International Space Station. Within this video, two distinctly-colored Astrobee robots [3] (a *yellow* robot and a *purple* robot) were observed enacting a fictitious maintenance and survey task[1]. In order to establish baseline awareness of the identities typically associated with these robot bodies, the two robots introduced themselves to the viewer, with the purple robot introducing itself as *Bumble* and the yellow robot introducing itself as *Honey*[2]. To help participants identify which robot body and identity was speaking, each body was animated to perform extralinguistic speech-accompanying movements (a "nodding" motion), and each identity was given a uniquely pitched voice.

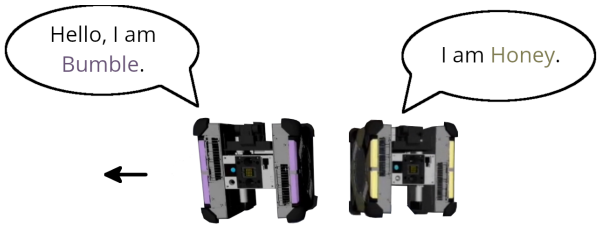

Figure 1: Astrobee Introduction

After this introduction, *Bumble*'s voice spoke through the purple body (with which it was originally associated) to state that it needed to perform a routine inspection of the station, after which the purple body was shown leaving the room, and the screen faded to black. The video then faded back onto the scene (now containing only the yellow body), and one of four video clips was shown, depending on the participant's experimental condition. Specifically, the participant's condition dictated the content and the delivery of a message conveyed to them through the yellow body.

[1]The mechanomorphic Astrobee robots were used in this work as the overarching motivations of this NASA-funded work centers around inspection tasks in distributed teams of Astrobee robots.

[2]We used these names as they are the official names given by NASA to the two original Astrobee robots currently operating aboard the ISS.

(a) Dissoc./Damaging Scenario **(b) Assoc./Building Scenario**

(c) Dissoc./Building Scenario **(d) Assoc./Damaging Scenario**

Figure 2: Robot Utterances under each Communication and Action Policy

The content of the message was determined by the between-subjects *Action Policy*. In the praiseworthy *Trust-Building Action* condition, participants were informed that *Bumble* had found a leak, whereas in the blameworthy *Trust-damaging Action* condition, participants were informed that *Bumble* had caused a leak.

The delivery of the message was determined by the between-subjects *Communication Policy*. In the *Body-Identity Associating Language* condition, the voice associated with the *Honey* identity spoke from the yellow robot body to relay information about *Bumble*, i.e., "Bumble has caused a leak." or "Bumble has found a leak." – a strategy that maintained the illusion of *Bumble* being a distinct agent with clearly associated body and identity. In the *Body-Identity Dissociating Language* condition, the voice associated with the *Bumble* identity spoke from the yellow robot body to relay information about "itself", i.e., "This is *Bumble*. I have found a leak." or "This is *Bumble*. I have caused a leak." – a strategy that broke the illusion of *Bumble* being a distinct agent with clearly associated body and identity, by allowing *Bumble* to speak through (momentarily "possess") the body originally associated with *Honey*.

After viewing this video, participants answered a number of survey questions, as described in the next section, after which they completed the experiment and were provided with payment.

3.2 Measures

In order to assess our hypotheses, we collected the following subjective measures through post-experiment surveys. To assess our hypotheses, each participant completed the Reliability and Capability subscales of the Multi-Dimensional Measure of Trust Survey [36] four times: once for each of the two bodies and once for each of the two identities, with the order in which these four surveys were presented counterbalanced across participants. Specifically, we measured perceived trustworthiness of each robot identity by asking participants to complete the aforementioned MDMT surveys, prefaced by instructions in which participants were told to provide

responses that best described their feelings or impressions of *Honey* or of *Bumble*. We then measured perceived trustworthiness of each robot body by asking participants to complete these MDMT surveys, prefaced by instructions in which participants were told to provide responses that best described their feelings or impressions of "the robot in the image", followed by an image of either the purple or yellow robot body.

The MDMT Surveys are composed of 7-point Likert items (from "Strongly Disagree" to "Strongly Agree") but also provide an option for participants to select "Does not apply" for each item rather than providing a 1-7 response. Hypothesis **H1** was assessed by examining the differences in number of "Does Not Apply" options selected under each Communication and Action Policy. Hypothesis **H2** was assessed by examining the differences in perceived trustworthiness of each body and identity under each Communication Policy. Hypothesis **H3** was assessed using both measures.

3.3 Participants

210 participants were recruited from Mechanical Turk (124 male, 83 female, 3 N/A). Participants ranged from 18 to 71 years (M=36.4, SD=10.8). 52 participants were assigned to the *Body-Identity Associating Language* and *Trust-Building Action* conditions; 56 were assigned to the *Body-Identity Associating Language* and *Trust-Damaging Action* conditions; 53 were assigned to the *Body-Identity Dissociating Language* and *Trust-Building Action* conditions; and 49 were assigned to the *Body-Identity Dissociating Language* and *Trust-Damaging Action* conditions. Participants were paid $2.02 for completing the study.

3.4 Analysis

We analyzed our anonymized data (accessible via the Open Science Framework, at https://osf.io/w7xdk/) with the JASP software package for statistical analysis [32, 60], using the default settings as justified by Wagenmakers et al. [63]. Bayesian t-tests [25, 65] with Communication Policy and Action Policy as grouping variables were respectively used to assess Hypotheses **H1** and **H3**. Bayesian Repeated-Measures ANOVAs with Communication Policy and Action Policy as between-subjects factors were used to assess Hypothesis **H2**, followed by computation of inclusion Bayes factor based on matched models ("Baws Factors") [37] for each candidate main effect and interaction, indicating (in the form of a Bayes Factor [39, 49]) for that effect the evidence weight of all candidate models including that effect compared to the evidence weight of all candidate models not including that effect.

For Bayesian ANOVAs, when sufficient evidence was found in favor of a main effect or when this evidence was inconclusive, the results were further analyzed using a post-hoc Bayesian t-test [25, 65] with a default Cauchy prior (center=0, r= $\sqrt{2}/2 = 0.707$). Finally, in this paper we follow the recommendations from previous researchers on linguistic interpretations of reported Bayes factors (BFs) [24].

While the Bayesian statistical approach has become widely used in the Cognitive Science and Psychology communities, it is still rare in the HRI community, and as such we will briefly describe the benefits of this approach. First, the use of a Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem). Second, the Bayesian approach allows investigators to examine the evidence both for and against hypotheses (whereas the frequentest approach can only quantify evidence towards rejection of the null hypothesis) [24]. Third, the Bayesian approach does not require reliance on p-values used in Null Hypothesis Significance Testing (NHST) which have recently come under considerable scrutiny [2, 54, 56, 62]. Finally, the Bayesian framework facilitates the use of previous study results to construct informative priors so that experiments may build upon the results of previous experiments rather than starting anew [34, 61].

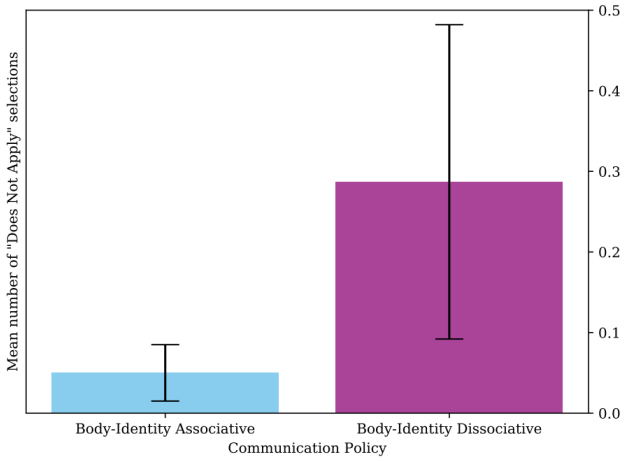

Figure 3: Weakness of Perceived Locus of Capability-based Trust in the Yellow Body (BF=4.9). In this and all other figures, error bars represent Standard Deviations.

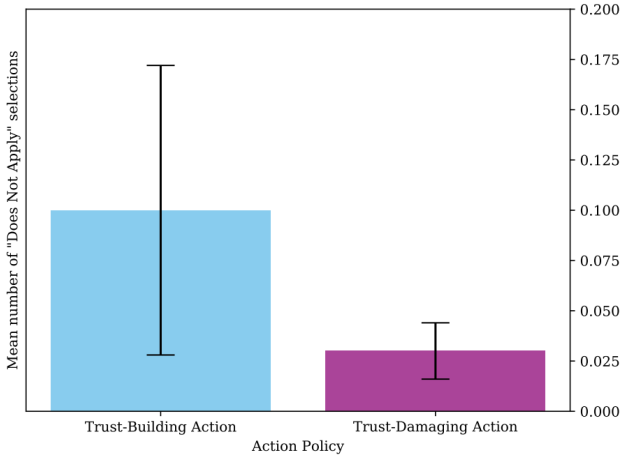

Figure 4: Weakness of Perceived Locus of Capability-based Trust in the *Bumble* Identity (BF=1.66).

4 RESULTS

In this section we present our experimental results, organized according to our three central hypotheses.

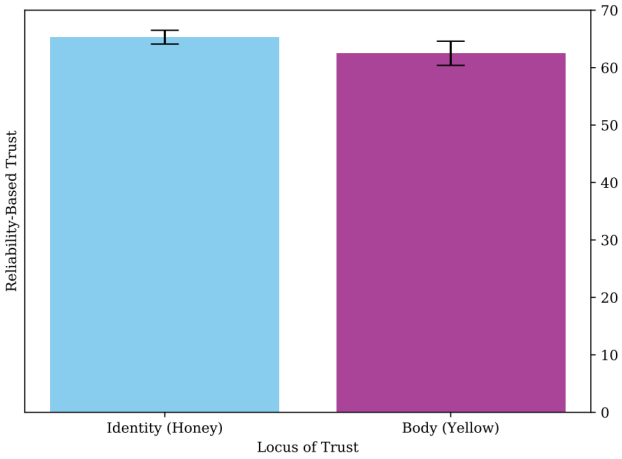

Figure 5: Divergence in Perceptions of Reliability-based Trustworthiness of Bodies and Identities (BF=10.231).

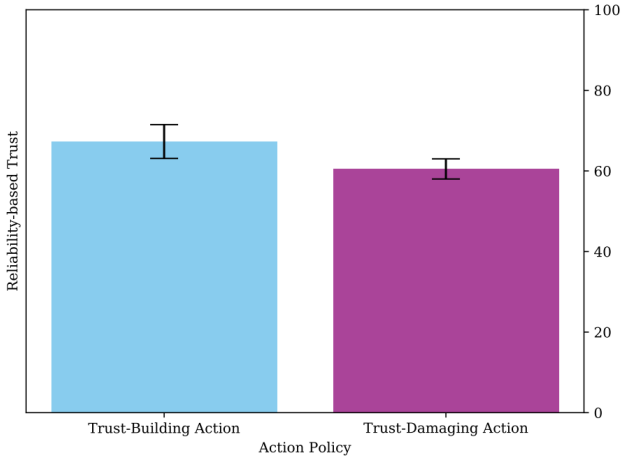

Figure 6: Effect of Action Policy on Perceived Reliability-based Trustworthiness in Yellow Body / *Honey* Identity Overall (BF=1.211).

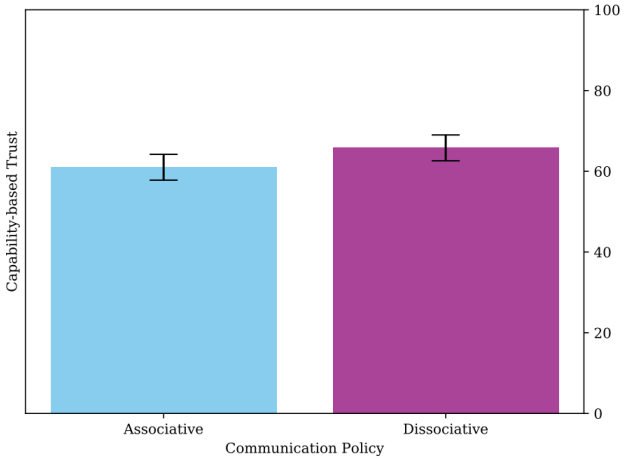

Figure 7: Effect of Communication Policy on Perceived Capability-based Trustworthiness in Purple Body / *Bumble* Identity Overall (BF=1.022).

4.1 Communication Policy and the Perceived Locus of Trust

Our first analysis consisted of a set of Bayesian t-tests with Communication Policy as a grouping factor, assessing the hypothesis that there would be a positive effect of body-identity dissociating communication (relative to body-identity associating communication) on the number of MDMT questions where participants selected "Does not apply", i.e., $H_+ : \delta > 0$ (i.e., a decrease in the strength of perceived locus of trust). Specifically, t-tests were used for both Reliability and capability-based trust, for both robot bodies, and for both robot identities.

As shown in Fig. 3, Bayes Factor analysis of the t-test results indicates evidence for H_+ only in the case of assessment of capability-based trust in the yellow robot body (BF=4.877), which means that our data are approximately 4.9 times more likely to occur under H_+ than under H_0, indicating moderate evidence in favor of H_+. A robustness analysis showed this result to be relatively stable across prior widths, ranging from about 2.846 to 6.851.

For the other analyses, the Bayes Factor analysis revealed negative or inconclusive evidence: Evidence was found against the hypothesized effect on reliability-based trust in the *Bumble* identity or its associated purple body, allowing such effects to be ruled out ($BFs < 0.333$). Anecdotal evidence was found in favor of a similar effect for capability-based trust in the *Honey* identity (BF=1.355), and anecdotal evidence was found against an effect of capability-based trust in capability-based trust in the *Bumble* identity (BF=0.422), reliability-based trust in the *Honey* identity (BF=0.527), capability-based trust in the purple body (BF=0.545), and reliability-based trust in the yellow body (BF=0.603), indicating these effects cannot be supported nor refuted until more data is gathered.

4.2 Divergence of Perceived Trustworthiness of Bodies and Identities

Our second analysis consisted of a set of Bayesian RM-ANOVAs with Communication Policy and Action Policy as between-subjects factors, and robot body and identity as within-subjects factors, assessing the hypotheses that there would be an interaction effects between Communication Policy and Locus of Trust (i.e., Body vs Identity) on human-robot trust, and between Action Policy and Locus of Trust on human-robot trust. Specifically, four Bayesian RM-ANOVAs were performed, for reliability and capability-based trust in the yellow body vs *Honey* identity, and the purple body vs the *Bumble* identity. We present our results separately for the yellow robot / *Honey* identity and purple robot / *Bumble* identity.

4.2.1 Yellow Body / Honey Identity. The Bayesian ANOVAs and subsequent Baws Factor Analysis for the Yellow Body / *Honey* Identity indicated strong evidence for an effect of Locus of Trust on reliability-based trust (BF=10.231), as shown in Fig. 5, suggesting

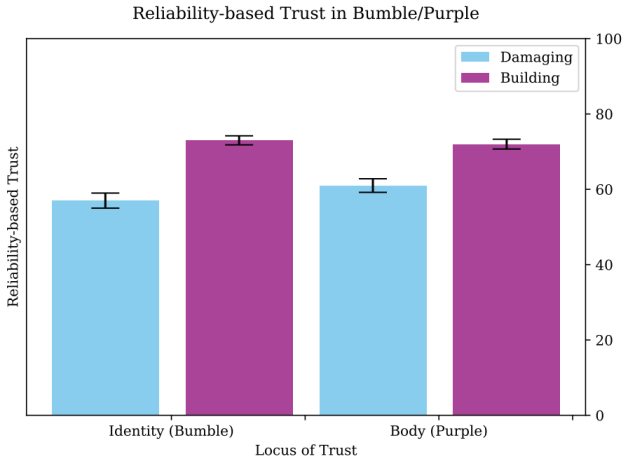

Figure 8: Effect of Action Policy on Divergence in Perceptions of Reliability-based Trustworthiness of Purple Body and *Bumble* Identity (BF=5.322).

that different levels of body- and identity-based reliability-based trustworthiness were perceived for the yellow robot / *Honey* identity, regardless of experimental condition. Anecdotal evidence was found in favor of an effect of Action Policy on reliability-based trust (BF=1.211), as shown in Fig. 6. Anecdotal evidence was found against an effect of Communication Policy on reliability-based trust (BF=0.365), against an effect of Action Policy on capability-based trust (BF=0.843), and against an interaction effect between Action Policy and Communication Policy on reliability-based trust (BF=0.569) or capability-based trust (BF=0.601). These inconclusive results suggest that these effects may exist, but more data is needed before they can be confirmed or refuted. Moderate evidence (BFs between 0.169 and 0.21) was found against all other effects.

4.2.2 Purple Body / Bumble Identity. The Bayesian ANOVAs and subsequent Baws Factor Analysis for the Purple Body / *Bumble* Identity indicated extreme evidence for an effect of Action Policy on both reliability-based trust (BF=3671) and capability-based trust (BF=4384) in the purple robot body and the *Bumble* identity, as trivially expected. Moderate evidence was also indicated for an interaction between Locus of Trust and Action Policy on reliability-based trust (BF=5.322), as shown in Fig. 8. Anecdotal evidence was found in favor of an effect of Communication Policy on capability-based trust (BF=1.022), as shown in Fig. 7. Anecdotal evidence was found against an effect of Communication Policy on reliability-based trust (BF=0.863), against interactions between Action Policy and Locus of Trust (BF=0.415) and Communication Policy (BF=0.524) on reliability-based trust, against differences in perceptions of trustworthiness of body and identity (regardless of experimental condition) on capability-based trust (0.650), and against an interaction between Communication Policy and Action Policy on capability-based trust (0.491). These inconclusive results suggest that these effects may exist, but that more data is needed before they can be confirmed or refuted. Moderate evidence (BFs between 0.135 and 0.201) was found against all other effects.

4.3 Action Policy and Perceived Locus of Trust

Our third analysis consisted of a set of Bayesian t-tests with Action Policy as a grouping factor, assessing the hypothesis that blameworthy actions would lead to a smaller number (relative to praiseworthy (trust-building) actions) of MDMT questions for which participants selected "Does not apply", i.e., $H_+ : \delta < 0$ (i.e., an increase in the strength of perceived locus of trust). Specifically, t-tests were used for both reliability and capability-based trust, for both robot bodies, and for both robot identities.

The Bayes Factor analysis revealed negative or inconclusive evidence for all cases: Evidence was found against reliability-based trust for the yellow robot body and *Honey* identity and for capability-based trust for the yellow robot body ($BFs < 0.333$). Anecdotal evidence was found against an effect of reliability-based trust in the purple robot body (BF=0.451) and the *Bumble* identity (BF=0.561), and for capability-based trust in the purple robot body (BF=0.4493) and the *Honey* identity (BF=0.337). Finally, as shown in the descriptive plot displayed in Fig. 4, anecdotal evidence was found in favor of an effect of capability-based trust in the *Bumble* identity (BF=1.66), suggesting that such an effect may exist, but that these effects cannot be supported nor refuted until more data is gathered.

5 DISCUSSION

We now discuss the implications of our results, as guided by our three research hypotheses.

H1: Do body-identity dissociating communication policies weaken trust loci?

Our results partially support our first hypothesis, as they provide moderate evidence that under the dissociative communication policy participants were less likely to view the yellow robot as a locus of capability trust. These results align with the two primary commitments of Deconstructed Trustee Theory, demonstrating the need to distinguish between different trust loci as well as the types of trust placed in those loci.

Accordingly, Deconstructed Trustee Theory allows us in this case to articulate the following general principle supported by our experimental results: *Robots that cede control of their bodies also cede their potential for capability trust.*

This suggests the following design guideline: *Designers who need to establish the capabilities of particular robot bodies should not allow the ostensible control of those bodies to appear to be involuntarily ceded to identities normally associated with other bodies.*

H2: Does body-identity dissociating communication lead to divergence between perceived trustworthiness of body and identity?

Our results did not support our second hypothesis, providing anecdotal to moderate evidence *against* an effect of communication policy on divergence between perceived trustworthiness of body and identity. However, our results provided evidence for a number of other relevant effects. Specifically, our results provided strong evidence suggesting that more reliability-based trust was built in the *Honey* identity than in the yellow body, regardless of Communication or Action policy. While it is not yet clear how to interpret this result, it nevertheless aligns with the two primary commitments of Deconstructed Trustee theory, demonstrating the need to

distinguish between different trust loci as well as the types of trust placed in those loci. Our results also suggest that body-identity dissociating communication policies may have led to increased perceived trustworthiness of the purple body and *Bumble* identity, perhaps due to increased communication "with" that identity.

H3: Do blameworthy actions strengthen trust loci?

Our results neither supported nor refuted the first clause of our third hypothesis: while an anecdotal effect was found suggesting that the locus of capability-based trust may have been strengthened by blameworthy actions for the *Bumble* identity, the evidence was not strong enough to accept this finding conclusively. We will note here that our chosen Bayes Factor threshold of 3.0 may be too high. Some statisticians have recently used simulations to show that when the null hypothesis is true, a BF as low as 1.2 yields only a 5% false-positive rate and a 10% false-negative rate [30]. In fact, if Frequentest statistics had been used in this experiment, a significant effect (assuming $\alpha = .05$) would have been found. However, a Frequentest analysis would not have afforded a flexible sampling plan as used in this work.

However, our results supported the second clause of this hypothesis: blameworthy actions led to increased divergence between perceived trustworthiness of body and identity, with lower perceived trustworthiness of the *Bumble* identity than in the purple body, and with the difference between these losses being more pronounced than trust gains between body and identity in response to praiseworthy actions. This result aligns with the two primary commitments of Deconstructed Trustee Theory, demonstrating the need to distinguish between different trust loci as well as the types of trust placed in those loci.

Accordingly, Deconstructed Trustee Theory allows us in this case to articulate the following general principle supported by our experimental results: *When a robot performs a blameworthy action, humans' moral cognitive processes lead them to identify which aspect of that robot (body or identity) is to blame for that action, resulting in greater trust losses for that locus.*

Similarly, our results provided interesting albeit inconclusive results regarding carry-over of the impact of blameworthy actions between robots. While perceived trustworthiness of the purple robot and *Bumble* identity were obviously negatively impacted by the purple robot taking blameworthy actions, our results also suggest that capability-based trust in the yellow robot and *Honey* identity may also have been impacted by these actions, even though it was not responsible for those actions. While the results do not conclusively support or rule out an effect, they suggest that the similarities between the two robots may well have led users to make negative inferences about one robot when the other made critical mistakes. This finding may well be worth examining through the lens of recent work in robot group entitativity [15].

6 CONCLUSIONS

In this paper, we introduced *Deconstructed Trustee Theory*, a new theory of human-robot trust which (1) stipulates that it is important not only to consider what type of trust is placed in a robot, but also to consider where that trust is placed, i.e., in which body- or identity-oriented trust locus, and (2) predicts (a) that different levels of trust can be accumulated in these loci, (b) that this body-identity trust dissociation may be effected by *body-identity dissociative* communication policies that reveal the potential for phenomena such as re-embodiment, co-embodiment, and agent migration in multi-robot systems, and (c) that blameworthy actions may trigger moral cognitive processes that lead to divergence between perceived trustworthiness of body and identity.

Our results provide support for Deconstructed Trustee Theory and for its first and second predictions. While no evidence was found for the ability of body-identity dissociating communication policies to lead to divergence between perceived trustworthiness of body and identity, our results in fact showed that this divergence occurs even without the use of such communication policies. Moreover, this work allowed us to demonstrate the performative nature of robot identities in distributed, integrated multi-robot systems, revealing a critical new dimension of robot interaction design.

Future Work

In future work, it will be critical for HRI researchers to further interrogate the new theories and concepts presented in this paper; probing the limits and implications of Deconstructed Trustee Theory and exploring the design space revealed by our dismissal of the assumption that robots must perform a tight 1-1 association between body and identity.

Future work should also explore the variety of strategies that may be used to communicate distinct identities, and the implications of those strategies. One limitation of the presented study is that the two selected voices varied in presented gender, with one sounding slightly more stereotypically female-presenting, and the other sounding slightly more stereotypically male-presenting. Although we do not expect this to have affected the results of this particular experiment, gender is clearly a central feature of how humans define and perform their identities, and there is evidence that human gender stereotypes and norms carry over into human-robot communication, mediating human perceptions of robots communication strategies [23]. As such, it will be critical to explore in future work what other design techniques may be employed to distinctly communicate the identity currently (albeit potentially temporarily) inhabiting or controlling a robot body, beyond gender-laden cues such as voice pitch.

Our work is also limited by our examination of Capacity Trust alone. While our nonsocial, task-oriented use context led us to specifically examine Capacity Trust in this work, our elevation of moral cognitive processes in this work suggests that future work might greatly benefit from similar investigation of Moral trust using the same methodology.

Finally, once campuses reopen [14] it will be critical to replicate this work through in-person experiments that enable deep immersion, human-robot rapport building, extended interaction, and the use of follow-up interviews to probe humans' mental models of their robot teammates.

ACKNOWLEDGMENTS

This work was supported by NASA Early Career Faculty award 80NSSC20K0070. We thank Trey Smith for his helpful feedback, Poulomi Pal for her undergraduate mentorship, and Shania Jo Runningrabbit for her further assistance.

REFERENCES

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

[2] James O Berger and Thomas Sellke. 1987. Testing a Point Null Hypothesis: The Irreconcilability of p-values and Evidence. *Journal of the American Statistical Association (ASA)* 82, 397 (1987).

[3] Maria G Bualat, Trey Smith, Ernest E Smith, Terrence Fong, and DW Wheeler. 2018. Astrobee: A New Tool for ISS Operations. In *2018 SpaceOps Conference*.

[4] Elizabeth J Carter, Samantha Reig, Xiang Zhi Tan, Gierad Laput, Stephanie Rosenthal, and Aaron Steinfeld. 2020. Death of a Robot: Social Media Reactions and Language Usage when a Robot Stops Operating. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 589–597.

[5] David Danks. 2019. The value of trustworthy AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 521–522.

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

[7] Daniel C. Dennett. 1989. *The intentional stance*. MIT press.

[8] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco. 2009. Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of the 8th ACM/IEEE international conference on Human-Robot Interaction (HRI)*.

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

[10] Timothy C Earle and George Cvetkovich. 1995. *Social trust: Toward a cosmopolitan society*. Greenwood Publishing Group.

[11] Chad Edwards, Autumn Edwards, Patric R Spence, and David Westerman. 2016. Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies* 67, 2 (2016), 227–238.

[12] Mica R Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors* 59, 1 (2017), 5–27.

[13] Anthony M Evans and Joachim I Krueger. 2009. The psychology (and economics) of trust. *Social and Personality Psychology Compass* 3, 6 (2009), 1003–1017.

[14] David Feil-Seifer, Kerstin S Haring, Silvia Rossi, Alan R Wagner, and Tom Williams. 2020. Where to next? The impact of COVID-19 on human-robot interaction research. *ACM Transactions on Human-Robot Interaction (T-HRI)* (2020).

[15] 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.

[16] Ilaria Gaudiello, Sébastien Lefort, and Elisabetta Zibetti. 2015. The ontological and functional status of robots: How firm our representations are? *Computers in Human Behavior* 50 (2015), 259–273.

[17] Mark Granovetter. 1985. Economic action and social structure: The problem of embeddedness. *American journal of sociology* 91, 3 (1985), 481–510.

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

[19] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. 2016. psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods* 48, 3 (2016), 829–842.

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

[21] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[22] Michita Imai, Tetsuo Ono, and Tameyuki Etani. 1999. Agent migration: communications between a human and robot. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, Vol. 4. IEEE, 1044–1048.

[23] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 559–567.

[24] Andrew F Jarosz and Jennifer Wiley. 2014. What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving* 7, 1 (2014), 2.

[25] Harold Jeffreys. 1938. Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* (1938), 161–198.

[26] Peter H Kahn, Aimee L Reichert, Heather E Gary, Takayuki Kanda, Hiroshi Ishiguro, Solace Shen, Jolina H Ruckert, and Brian Gill. 2011. The new ontological category hypothesis in human-robot interaction. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 159–160.

[27] 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. In *Thirteenth International Conference on Human=Computer Interaction*.

[28] Michael Kriegel, Ruth Aylett, Kheng Lee Koay, K Casse, Kerstin Dautenhahn, Pedro Cuba, and Krzysztof Arent. 2010. Digital body hopping-migrating artificial companions. *Proceedings of Digital FuturesâĂŹ 10* (2010).

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

[30] Daniel Lakens. 2016. Power analysis for default Bayesian t-tests. (2016). http://daniellakens.blogspot.com/2016/01/power-analysis-for-default-bayesian-t.html

[31] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[32] Jonathon Love, Ravi Selker, Maarten Marsman, Tahira Jamil, Damian Dropmann, Josine Verhagen, Alexander Ly, Quentin F Gronau, Martin Smira, Sacha Epskamp, et al. 2019. JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software* 88, 2 (2019), 1–17.

[33] 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.

[34] Alexander Ly, Alexander Etz, Maarten Marsman, and Eric-Jan Wagenmakers. 2018. Replication Bayes factors from evidence updating. *Behavior Research Methods* (13 Aug 2018). https://doi.org/10.3758/s13428-018-1092-x

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

[36] Bertram F Malle and Daniel Ullman. 2020. A Multi-Dimensional Conception and Measure of Human-Robot Trust. (2020).

[37] Sebastiaan Mathôt. 2017. Bayes like a Baws: Interpreting Bayesian repeated measures in JASP. *cogsci.NL* (2017). https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp

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

[39] RD Morey and JN Rouder. 2015. BayesFactor (Version 0.9. 11-3)[Computer software]. (2015).

[40] Douglass North. 1990. Institutions, institutional change and economic performance Cambridge University Press. *New York* (1990).

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

[42] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G Jentsch. 2013. Building appropriate trust in human-robot teams. In *2013 AAAI Spring Symposium Series*.

[43] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.

[44] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[45] Samantha Reig, Michal Luria, Janet Z Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2020. Not Some Random Agent: Multi-person interaction with a personalizing service robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 289–297.

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

[47] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–108.

[48] Julian B Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of personality* (1967).

[49] Jeffrey N Rouder, Richard D Morey, Paul L Speckman, and Jordan M Province. 2012. Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology* 56, 5 (2012), 356–374.

[50] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review* 23, 3 (1998), 393–404.

[51] Kristin Schaefer. 2013. *The Perception and Measurement of Human-Robot Trust*. Ph.D. Dissertation. University of Central Florida.

[52] Lawrence Shapiro. 2010. *Embodied cognition*. Routledge.

[53] Blair H Sheppard and Dana M Sherman. 1998. The grammars of trust: A model and general implications. *Academy of management Review* 23, 3 (1998), 422–437.

[54] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 11 (2011).

[55] Nicolas Spatola, Sophie Monceau, and Ludovic Ferrand. 2019. Cognitive Impact of Social Robots: How Anthropomorphism Boosts Performances. *IEEE Robotics & Automation Magazine* (2019).

[56] Jonathan AC Sterne and George Davey Smith. 2001. Sifting the Evidence – What's Wrong with Significance Tests? *Physical Therapy* 81, 8 (2001), 1464–1469.

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

[58] Xiang Zhi Tan, Michal Luria, and Aaron Steinfeld. 2020. Defining Transfers Between Multiple Service Robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 465–467.

[59] Xiang Zhi Tan, Samantha Reig, Elizabeth J Carter, and Aaron Steinfeld. 2019. 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)*. IEEE, 114–122.

[60] JASP Team et al. 2018. JASP (Version 0.9. 0.1). *Computer software. Available at: https://jasp-stats. org* (2018).

[61] Josine Verhagen and Eric-Jan Wagenmakers. 2014. Bayesian Tests to Quantify the Result of a Replication Attempt. *Journal of Experimental Psychology: General* 143, 4 (2014), 1457–1475.

[62] Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review* 14, 5 (2007), 779–804.

[63] Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F Gronau, Damian Dropmann, Bruno Boutin, et al. 2018. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review* 25, 1 (2018), 58–76.

[64] Samantha F Warta, Katelynn A Kapalo, Andrew Best, and Stephen M Fiore. 2016. Similarity, complementarity, and agency in HRI: Theoretical issues in shifting the perception of robots from tools to teammates. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 1230–1234.

[65] Peter H Westfall, Wesley O Johnson, and Jessica M Utts. 1997. A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 84, 2 (1997), 419–427.

[66] 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* 4, 2 (2015), 24–49.

[67] Oliver E Williamson. 1993. Calculativeness, trust, and economic organization. *The journal of law and economics* 36, 1, Part 2 (1993), 453–486.

[68] Lynne G Zucker. 1986. Production of trust: Institutional sources of economic structure, 1840-1920. *Research in organizational behavior* 8 (1986), 53–111.