A Gender-Controlled Replication Study on Deconstructed Trustee Theory

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

Understanding what influences human trust in robots is critical to avoid mistrust, distrust, and overtrust in human-robot interactions. However, the non-humanlike nature of robot identity in multi-robot systems can complicate how humans conceptualize robots and the nature of human-robot trust. Prior work by Williams et al. explored this complexity and introduced Deconstructed Trustee Theory, a theory on human-robot trust that accounts for these dynamics. However, confounding factors relating to robot gender identity may have impacted their findings. Thus, in this work, we sought to replicate their findings while controlling for robot gender identity through an online human-subject study (n=189). Through this replication, we verify key findings from prior work and provide further support for Deconstructed Trustee Theory.

KEYWORDS

Trust, Identity Performance, Robo-Identity, Multi-Robot Systems

1 MOTIVATION

Trust plays a crucial role in human-robot interaction (HRI) [6, 7]. It is essential for human trust in robots to be well-calibrated to avoid the misuse of robots which otherwise may cause harm [5]. As such, it is important to understand what influences human trust in robots and how well-calibrated trust in robots can be built and maintained. However, the nature of robots and how humans conceptualize robots can complicate how and where trust is built and maintained, especially when accounting for robot identity in multi-robot systems [2, 11].

Robots do not have to maintain a tight, humanlike association between physical body and performed identity or persona. Instead, robots may leverage different identity performance strategies to change how robot identity is presented and distributed across different robot bodies. Identity performance strategies can be used to present non-humanlike body-identity associations such as re-embodiment, in which an identity can shift its presence from one robot body to another [9, 10]. This strategy can result in the transfer of trust across embodiments [8], and can lead to greater trust resilience in the face of robot failures [9]. Moreover, identity performance strategies can fundamentally change the ways humans mentally model groups of robots and determine whether and how to trust them [1, 2]. However, the flexibility of robot identity raises key challenges in who or what is being trusted when trust is assessed in “a robot” – is it the robot’s body, or the robot’s identity?

This question is addressed through Williams et al. [11]’s *Deconstructed Trustee Theory*, which argues for differential levels and concreteness of trust in robot bodies versus identities. However, we ask whether the evidence used to justify Deconstructed Trustee Theory may have been compromised by certain confounding factors, specifically, the ways that robot gender identity was performed in Williams et al. [11]’s work. In Williams et al. [11]’s experiment, robots used names that carried distinct implications for their perceived warmth and competence in part due to gendered expectations. This is concerning as human gender stereotypes and norms can carryover to HRI and affect human preferences regarding robot interactions [4] In this work, we thus present the results of an experiment (n=189) that replicates Williams et al. [11]’s methodology while controlling for robot gender identity. Our results generally support Deconstructed Trustee Theory, while highlighting which specific elements of the theory are most supported.

2 DECONSTRUCTED TRUSTEE THEORY

In prior work, Williams et al. [11] introduced *Deconstructed Trustee Theory* which argues that when we talk about human-robot trust, we should differentiate between different trust *loci*, such as trust in the robot’s *body* versus trust in its *identity*. This theory predicts that (1) different levels of trust may be built/lost in each trustee’s loci of trust, (2) different robot identity performance strategies may differently affect the trust built/lost in each loci of trust, and (3) different trust-affecting actions may differently affect the trust built/lost in each loci of trust.

To test these predictions, Williams et al. [11] conducted an online human-subjects study in which participants were shown a short video of two Astrobee robots [3] (each distinctly colored and named: a purple robot named “Bumble" and a yellow robot named “Honey"). The two robots were shown participating in a maintenance and survey task, with the yellow robot staying with the participant, and the purple robot flying away. At the end of the video, the yellow body was used to communicate the presence of an air leak to the participant. Williams et al. [11] varied the nature of this communication along two dimensions (Fig. 1). First, Williams et al. [11] varied the nature of the robots’ *Actions*. In the *Trust-Damaging* condition, Bumble, the identity in the purple body, *caused* the air leak. In the *Trust-Building* condition, Bumble *discovered* the air leak. Second, Williams et al. [11] varied *Communication*: in the *Body-Identity Associating Language*, the Honey voice came from the yellow body and relayed information about the trust building or damaging action from Bumble; in the *Body-Identity Dissociating Language* condition, the Bumble voice came from the yellow body, allowing the Bumble identity to share the information about the trust building or damaging action for itself, by temporarily “posessing” the yellow body. Williams et al. [11] assessed the commitments of Deconstructed Trustee Theory by asking participants in each condition to assess each body and identity using variants of the

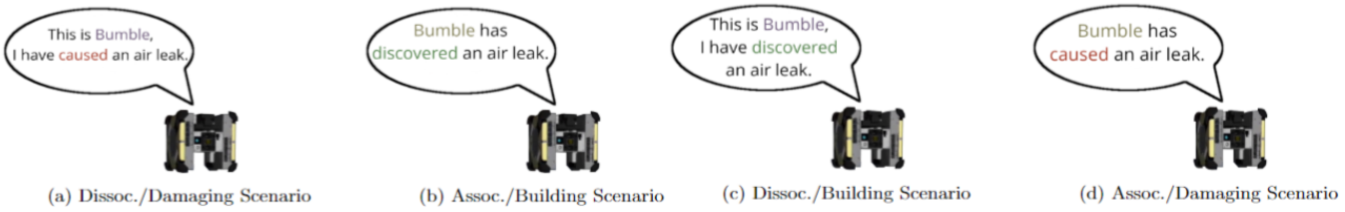

This is Bumble,
I have caused an air leak.
(a) Dissoc./Damaging Scenario
Bumble has discovered an air leak.
(b) Assoc./Building Scenario
This is Bumble,
I have discovered an air leak.
(c) Dissoc./Building Scenario
Bumble has caused an air leak.
(d) Assoc./Damaging Scenario

Figure 1: Study video conditions used by Williams et al. [11] to communicate presence of an "air leak".

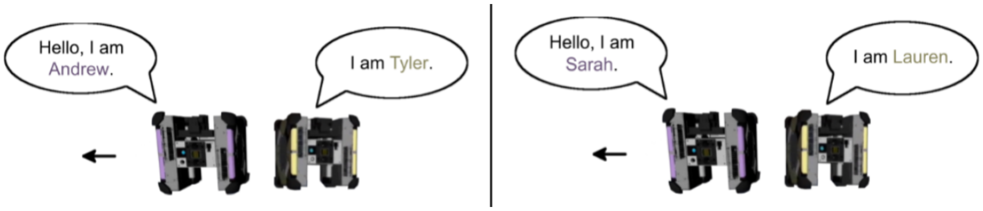


Figure 2: Introductions used in replication. Left side is the *Masculine Robot*, right side is the *Feminine Robot* condition.

reliability and capability trust subscales of the Multi-Dimensional Measure of Trust survey [7].

Overall, Williams et al. [11] found support for several key predictions of *Deconstructed Trustee Theory*: (1) The use of Body-Identity Dissociating Language made people less likely to view the yellow robot body as something it was more reasonable to consider trusting (at least for capability-based trust); (2) Different levels of trust were placed in the robot bodies and identities (at least for reliability-based trust in the yellow robot body, which was trusted less than the "Honey" identity inhabiting it); (3) Similarly, when trust damaging actions were taken, trust divergence increased for the purple body and "Bumble" identity, with greater trust losses for the identity than for the body.

These results thus supported the overall premise of Deconstructed Trustee Theory, although support was not found for all hypothesized effects, and although only weak evidence was found for certain effects. In addition, as Williams et al. [11] point out, it is possible that their results may have been compromised by the names used in their experiment. In their work, "Honey" and "Bumble" were used because those are names used by NASA for the Astro-"bees" aboard the ISS. Yet Honey may be more likely to be gendered as feminine, and may be a name from which increased warmth might be inferred. Similarly, Bumble may be more likely to be gendered as masculine, and may be a name from which decreased competence might be inferred. Indeed, robot voice and name are key identity performance cues used to convey and infer robot gender [12]. To develop better confidence in Williams et al. [11]'s results, we thus sought to replicate their work while controlling for the gender of the names used, and while removing the warmth and competence oriented connotations from those names.

3 METHOD

We conducted our replication experiment through the Prolific crowdsourcing platform (prolific.co). Our procedure was identical to that used by Williams et al. [11], with only the robot names and voices changed. Specifically, the robots in our experiment introduced themselves using intentionally gendered names and voices (Figure 2), according to experimental condition. In the *Masculine Robot* condition, the purple robot introduced itself as *Andrew* and the yellow robot introduced itself as *Tyler*, each using a masculine voice. In the *Feminine Robot* condition, the purple robot introduced itself as *Sarah* and the yellow robot introduced itself as *Lauren*, each using a feminine voice.

We recruited 189 participants (104 Male, 85 Female) who ranged from 18 to 64 years (M=28.6, SD=9.6). Participants were randomly assigned to conditions, resulting in 21-27 participants per experiment cell under our 2 (Communication) × 2 (Action) × 2 Robot Gender experimental design.

4 RESULTS

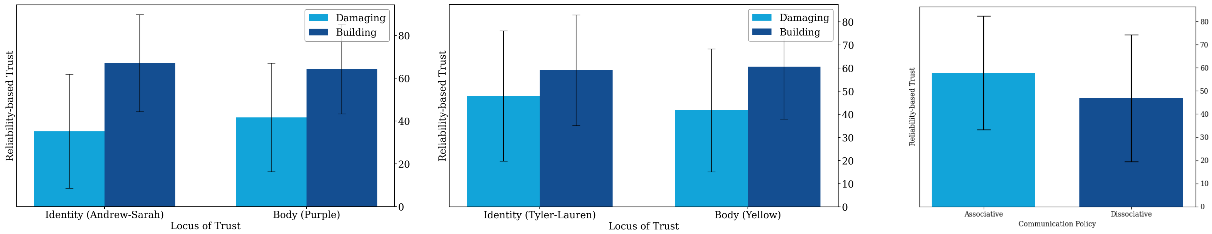
4.1 Effect of Robot Gender Identity

Before reporting the main results of this study, we briefly present the effects of robot gender. Two One-Sided T-Tests demonstrated equivalence between the Masculine Robot and Feminine Robot conditions, allowing us to remove this factor from our analysis and providing evidence that Williams et al. [11]'s results were not likely to have been impacted by this potential confound.

4.2 Replication Analysis

With this added variable removed, we proceeded by replicating Williams et al. [11]' statistical analysis.

We were not able to replicate Williams et al. [11]'s finding that under the body-identity dissociating language policy (in which the Bumble identity "possessed" the yellow body to deliver its message), participants viewed the yellow robot as a weaker locus for potential trust. However, we instead found that in this condition, both the yellow body and the associated identity suffered especially large trust losses (Figure 3(c)).



((a))

((b))

((c))

Figure 3: (a) Effect of Action Policy on Divergence in Perceptions of Reliability-based Trustworthiness of Purple Body and *Andrew-Sarah* Identity (BF=99.866). (b) Effect of Action Policy on Divergence in Perceptions of Reliability-based Trustworthiness of Yellow Body and *Tyler-Lauren* Identity (BF=40.389). (c) Effect of Communication Policy on Perceived Reliability-based Trustworthiness in Yellow Body / *Tyler-Lauren* Identity (BF=11.403). In all figures, error bars represent Standard Deviations.

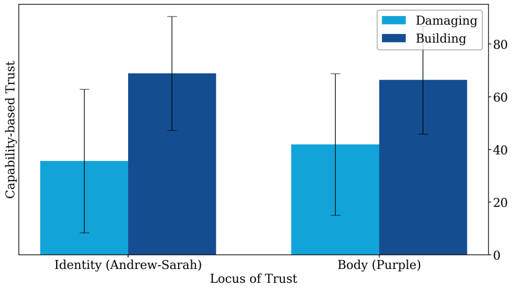


Figure 4: Effect of Action Policy on Divergence in Perceptions of Capability-based Trustworthiness of Purple Body and *Andrew-Sarah* Identity (BF=100.469).

Similarly, we did not find an overall difference in trust between the yellow body and its associated identity. However, we did replicate their finding that trust-damaging actions (by the purple body / Bumble identity) led to divergence in trust in the purple robot's body and it's associated identity, i.e., that when the purple body and its associated identity were reported to have *caused* an air leak, this led to greater loss of trust in the identity than in the purple body. Moreover, this effect was found for both reliability (Figure 3(a)) and capability trust (Figure 4), whereas in previous work the finding was limited to capability trust. In addition, we found that trust-damaging actions (by the purple body / Bumble identity) led to loss of trust in the yellow robot body, which was used to report those trust-damaging actions under both communication policies (Figure 3(b)), i.e., participants had a tendency to "shoot the messenger".

5 DISCUSSION AND CONCLUSION

In this work, we sought to replicate the findings by Williams et al. [11] while controlling for robot gender identity through an online human-subject study (n=189). Through this work, we were able to rule out the potential effects of robot gender identity on trust, and provided further support for Deconstructed Trustee Theory.

Overall, our findings provide a more nuanced understanding of the ways that the trust in a robot's body and identity can be affected to different extents by trust-damaging actions, with the greatest losses of trust going to the identity of damage-causing robots and the bodies of the "messengers" of that damage. Moreover, our results suggest that previous work may have been too quick to rule out effects of communication policy, as Body-Identity Dissociating Language lead to decreased perceptions of trust for both the reporting robot body (yellow) and its associated identity.

Both the prior study by Williams et al. [11] and our replication jointly demonstrate the importance of separately considering robot body and identity, and the roles that robot identity performance strategies and trust-affecting language may have on different types of trust. However, as both studies were done online where participants were only observing an interaction between robots, future work will need to further investigate these trust dynamics through in-person studies that more closely mirror actual human-robot interactions. Future work should also explore the variety of strategies that may be used to perform identities, and the implications of those strategies. One limitation of the presented studies, for example, is that only name and voice are used to perform identity, which may constrain the types of mental models formed by participants [1]. As such, it will be critical to explore other design cues in future work.

ACKNOWLEDGMENTS

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

Alexandra Bejarano is a PhD student at the Colorado School of Mines. **Tom Williams** is an Associate Professor of Computer Science at the Colorado School of Mines.

REFERENCES

[1] Alexandra Bejarano, Samantha Reig, Priyanka Senapati, and Tom Williams. 2022. You had me at hello: The impact of robot group presentation strategies on mental model formation. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 363–371.

[2] Alexandra Bejarano and Tom Williams. 2023. No Name, No Voice, Less Trust: Robot Group Identity Performance, Entitativity, and Trust Distribution. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1339–1346.

[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] 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.

[5] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1 (2020), 297–309.

[6] Theresa Law and Matthias Scheutz. 2021. Trust: Recent concepts and evaluations in human-robot interaction. *Trust in human-robot interaction* (2021), 27–57.

[7] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.

[8] Kohei Okuoka, Kouichi Enami, Mitsuhiko Kimoto, and Michita Imai. 2022. Multi-device trust transfer: Can trust be transferred among multiple devices? *Frontiers in Psychology* 13 (2022), 920844.

[9] Samantha Reig, Elizabeth J Carter, Terrence Fong, Jodi Forlizzi, and Aaron Steinfeld. 2021. Flailing, hailing, prevailing: perceptions of multi-robot failure recovery strategies. In *Int'l Conf. HRI*.

[10] Samantha Reig, Michal Luria, Elsa Forberger, Isabel Won, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2021. Social robots in service contexts: Exploring the rewards and risks of personalization and re-embodiment. In *Designing Interactive Systems Conference*.

[11] 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*. 262–271.

[12] Katie Winkle, Ryan Blake Jackson, Alexandra Bejarano, and Tom Williams. 2021. On the flexibility of robot social identity performance: benefits, ethical risks and open research questions for HRI. In *HRI Workshop on Robo-Identity*.