Emerging Technologies and Future of Work, Vol. 117, 2023, 88–97
https://doi.org/10.54941/ahfe1004408

Validating Trust in Human-Robot Interaction Through Virtual Reality: Comparing Embodied and "Behind-the-Screen" Interactions

Sebastian S. Rodriguez[1], Harsh Deep[1], Drshika Asher[1], James Schaffer[2], and Alex Kirlik[1]

[1] University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA
[2] Independent, Katy, TX, 77494, USA

ABSTRACT

Human-agent interaction is commonplace in our daily lives, manifesting in forms ranging from virtual assistants on websites to embodied agents like robots that we might encounter in a physical setting. Previous research has largely been focused on "behind-the-screen" interactions, but these might not fully encapsulate the nuanced responses humans exhibit towards physically embodied agents. To address this gap, we use virtual reality to examine how simulated physical embodiment and the reliability of an agent (automated robotic crane) influence trust and performance in a task simulating a quality assurance role and compare it to a "behind-the-screen" interaction. Out of 119 participants, the data revealed there is a marked behavioral shift observed when reliability hits a 91% threshold, with no influence from embodiment. Furthermore, participants displayed a tendency to trust and defer to the decisions of embodied agents more, especially when these agents were not infallible. This study accentuates the need for transparency about an agent's capabilities and emphasizes the significance of ensuring that the agent's representation is congruent with the nature of the interaction. Our findings pave the way for a deeper understanding of human-agent interactions, suggesting a future where these interactions might seamlessly blend the virtual and physical realms.

Keywords: Human-robot interaction, Embodied agents, Signal detection theory, Virtual reality, Trust calibration, Decision support systems

INTRODUCTION

Research in human-agent interaction often employs an interaction paradigm wherein participants interact with virtual agents through a screen, rendering the agents intangible. However, as computational advancements unfold, AI systems are likely to extend beyond virtual domains, boasting tangible embodiments that necessitate interactions with other entities, either human or agent. Robots, for instance, play vital roles in tasks ranging from assembly, packaging, space exploration, medical surgery, to mass production and safety. The domain of human-robot interaction delves deep into the dynamics of human engagement with physically represented robots, with a wide history of investigating trust and human responses to varied robot capabilities and appearances (Chien et al., 2018; Esterwood et al., 2021; Li, 2015; Natarajan & Gombolay, 2020; Tian et al., 2021; Tolmeijer et al., 2020; van den Brule et al., 2014). Yet, a significant gap persists in the research that measures trust through the manipulation of robot representation, tangible or otherwise, and even more so through the lens of reliability. Traditional experiments involving physical robots are often restricted by budgetary and safety concerns. Virtual reality (VR), on the other hand, has quickly risen to prominence for its ability to authentically simulate embodied physical interactions in a risk-free, high-fidelity setting — a methodology that has proved instrumental in academic research (Lee-Cultura & Giannakos, 2020; Saffo et al., 2020). Leveraging VR provides a comprehensive view of embodied interactions. We aim to address the gap with a comparative analysis of trust models between virtual and physical domains, especially in terms of how embodiment influences trust calibration and decision-making. Consequently, our inquiry revolves around the following research questions:

- How are embodied interactions decisions made and trust formed during collaborative decision-making tasks?
- In what manner does embodiment influence perception of agent capability across different reliability levels?

We conduct a mixed design with 119 participants to complete a quality assurance scenario in either a computer or in VR, and find that both embodiment and reliability affect what and how decisions are made when paired with an agent.

BACKGROUND

Embodied Agents and Their Impact on Trust

Within the domain of human-robot interactions, the role of embodied agents stands out prominently. An embodied agent is a physical entity, driven by computational logic that allows it to interact with the world, be it physically or virtually. These entities, colloquially termed as "bodies", are equipped with appendages, maybe heads or hands, which enable their active engagement with their immediate environment via sensors and motors (Tonkin et al., 2017; Ziemke, 2001). Notably, these "bodies" do not require a human-like representation; they can embody mechanoid or creature-like designs (Wang et al., 2018). Users' interactions with embodied agents significantly differ from those with non-embodied counterparts (Hertzum et al., 2002). Such embodiments influence users' trust (Rae & Takayama, 2013), empathy (Seo et al., 2015), and attention (Wainer et al., 2006). Humans tend to perceive humanoid or anthropomorphized agents as more trustworthy and diligent (Lawson-Guidigbe et al., 2020; Walters et al., 2009; Wang et al., 2018). This leads to enhanced trust dynamics, including the ability to repair trust after a malfunction or loss of performance (de Visser et al., 2016). Consequently, system designers often resort to incarnating their agents as user interfaces, manifesting as avatars, chatbots, or recommendation systems.

Prior research widely suggests that the effectiveness of agent features, such as reliability and transparency, shifts based on the context of their deployment. Yet, research yields inconclusive evidence on embodiment's influence over trust in human-agent collaborations. Herse et al., for instance, discerned no embodiment impact in high-risk tasks under pressing timelines (Herse et al., 2018). Conversely, Mollahosseini et al. contended that embodiment enhanced the discernibility of robots' facial cues, albeit restricted to certain emotions (Mollahosseini et al., 2018). Such disparities suggest that the effectiveness of embodiment largely hinges on the context it is deployed in.

Tangible vs. Intangible Agents

The level of tangibility granted to embodied agents is notably diverse (Schaffer et al., 2020). Tangible agents possess a tactile body, such as robots, drones, or even physical avatars of virtual assistants like Amazon Echo. Their intangible counterparts, however, are tethered to the virtual domain, but might visually emulate tangible entities. Interactions with tangible agents are uniquely shaped by social cues, cultural traditions, system expectations, and varying degrees of acceptance towards anthropomorphic designs (Breazeal, 2004). For instance, tangible robot swarms incite distinct psychophysiological human responses compared to a virtual simulation representing the same (Podevijn et al., 2016). The dynamics of Human-Agent Teams are acutely receptive to tangibility nuances. Although initial interactions suggest heightened trust and politeness towards tangible entities, sustained interactions might dilute this trust (Kulms & Kopp, 2016). Intangible agents, being economically efficient and widely accessible, persist as significant subjects of study, holding promise for specific implementations.

Overall, our discussions converge on a signal detection theory (SDT) task, orchestrated in the company of an embodied robot. By leveraging VR, we aim to assess the interplay between reliability in decision-making and agent embodiment, by using simulated tangibility. This endeavor seeks to discern if reliability impacts, as observed in prior studies, extend to tangible realms, and to ascertain the efficacy of virtual reality as a surrogate for tangible human-robot interaction research.

METHODOLOGY

Simulation Design

We designed a collaborative decision-making scenario in the Unity game engine, dubbed *Warehouse*. Participants are tasked with the role of a quality assurance worker, who needs to ensure that several orders shipping from warehouse have the correct item packed. If the package's contents match the order, they should send the order, and if they do not, they should reject it – following an SDT-based task. However, two main constraints are presented. First, the package cannot be opened to check its contents. Instead, workers are provided with a package scanner that reveals the content inside. Unfortunately, the scanners are unreliable, and present with a high amount of visual interference. To assist with decision-making, every worker is partnered with a robotic teammate, that can determine the package's contents and provide a recommendation on whether a package should be sent or rejected. To test for reliability, the accuracy of the recommendations will vary according to condition. Second, worker performance is rated based on how quickly packages are correctly processed. Good performance is rewarded with a monetary bonus at the end of the simulation.

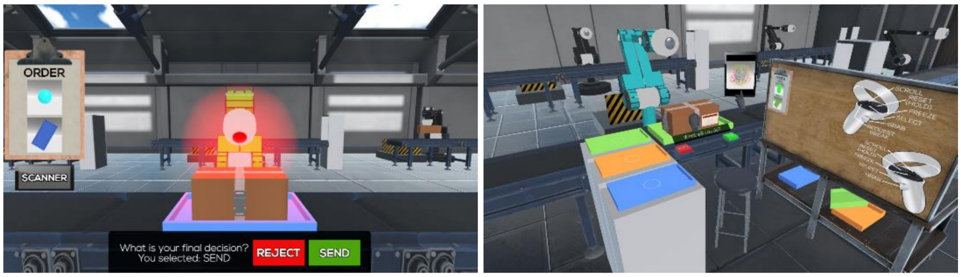
What is your final decision?
You selected: SEND

Figure 1: Two screenshots of the *Warehouse* scenario. (Left) presents the scenario as a computer game interactable with a mouse and keyboard. (Right) presents a sideview of the VR environment, where participants can grab and interact with the objects.

Depending on condition, participants will either complete the scenario using a computer monitor with mouse input (Screen Representation condition) or in virtual reality using a Meta Quest 2 headset (VR Representation condition). The Meta Quest 2 includes handheld motion-tracked controllers, which allows for embodied interactions with the virtual environment when processing orders during the simulation. As part of the scenario, participants are first instructed in the procedure to process an order, in the form of onboarding videos. First, an order is received on the clipboard (an interface popup in the Screen condition or a virtual object in the VR condition); this order contains the item that must be sent, along with the colored tray that must be placed before the robot brings the package from storage. Once the correct tray is placed, the robot places the package in front of the participant and waits for the participant's initial decision. Once a decision has been made, the robot states its recommendation in the form of an alarm or lack thereof. When an alarm is sounded, it is a direct cue that it detected the item inside the package does not match the order. Participants are then given the opportunity to change their decision considering the robot's recommendation. This procedure is repeated for all orders.

As mentioned, we introduce time pressure to place an ongoing urgency, much like real-world workers must fulfill quotas. At every order, participants can earn points if they complete an order correctly (true cases – true positives and false positives) or lose points if they fail to do so (false cases – false positives and false negatives). Initially, the participant can receive a maximum reward or a minimal penalty, respectively. As time progresses, the reward and penalty are linearly scaled accordingly to de-incentivize participants from delaying their decision, minimizing rewards and maximizing penalties at 10 seconds (selected through pilot testing). We expect that participants will use and adhere to the robot's recommendation in their search to be speedy workers (Rice et al., 2008), although they may still choose to rely on their own visual acuity for their decisions. This reward structure then translates into a bonus for their given compensation; prior research has shown this to be apt at motivating participants to make sound and appropriate decisions (Bansal et al., 2019; Zhang et al., 2020). Table 1 outlines the distribution of rewards and penalties per trial during the *Warehouse* simulation.

Table 1. Reward matrix for the *Warehouse* simulation.

		Recommendation			
		Send		**Reject**	
Package	Match	TP		FN	
		$t = 0$ +5	$t = 10$ +1	$t = 0$ 0	$t = 10$ −2
	Mismatch	FP		TN	
		$t = 0$ −1	$t = 10$ −5	$t = 0$ +2	$t = 10$ 0

Experimental Design

For this study, we study two independent variables as discussed earlier: Representation and Reliability. Representation describes whether the interaction and the agent are embodied, divided in 2 levels: Screen and VR. Reliability then describes the accuracy of recommendations participants receive from the robotic teammate: 100% (Perfect), 91% (Ideal), 75% (Good Enough), 50% (No Info). The Representation condition was presented between-subjects, whereas the Reliability condition was presented within-subjects. This results in a 4 block study with 24 orders during each block. For each block, participants complete 12 orders, receive an accuracy and speed report, and complete the next 12 orders. The sequence and errors of the orders were standardized across participants to maintain consistency and control trust recovery rates. The Reliability blocks were counterbalanced using a randomized balanced Latin square design.

The dependent variables we focused on are decision-making choices, number of deferrals, and level of trust calibration. Overall performance is measured by the amount of points earned across all 4 blocks, as described by the reward matrix in Table 1. Decision-making was measured by the amount of reliance and compliance demonstrated during the orders. Reliance describes if a package was sent when the robotic teammate gave no alarm (i.e., no action given lack of a cue), whereas compliance describes if a package was rejected when the robotic teammate gave an alarm (i.e., action when an alarm is given). We inferred deferral from a combination of 2 explicit behaviors: for a given order, if participants used the scanner and fulfilled the order in a combined time less than 2 seconds. This behavior could be construed as automatically adhering to the robot's recommendation.

Participants were recruited from Prolific, an online platform chosen for its high-quality data compared to other crowdsourcing platforms, especially during the challenges of in-person recruiting due to COVID-19. After screening for colorblindness, corrected vision, and hearing impairments, approved participants filled out a demographic survey and pre-survey metrics, then installed and familiarized themselves with the *Warehouse* scenario through instructional videos (where the robotic teammate was framed as imperfect) and an attention-check quiz. They underwent 10 practice trials without robot recommendations before tackling four blocks of 24 trials, each having distinct robot reliability and emphasizing different calibration settings.

RESULTS

To establish patterns of behavioral trust, we focus on reliant and compliant behavior when interacting with the robotic teammate. Both reliance and compliance have been found to relate to over-trusting and complacent behavior (Dixon & Wickens, 2006). For reliance, the ANOVA revealed a main effect in the Reliability (F(3, 313) = 11.12, p < 0.001) and Time (F(1, 117) = 5.59, p < 0.05) factors, with an interaction between them (F(3, 351) = 45.18, p < 0.001). Opposite to compliance, after the mid-block feedback, reliance decreased in the No Info condition, where it increased in the Ideal condition. Other reliability conditions remained unchanged. For compliance, the ANOVA revealed a main effect in the Reliability factor (F(3, 351) = 17.47, p < 0.001), with an interaction between Reliability and Time (F(3, 351) = 69.03, p < 0.001). Compliance was correlated with reliability. After the mid-block feedback, compliance increased in the No Info condition, whereas it decreased in the Ideal condition. Other reliability conditions remained unchanged.

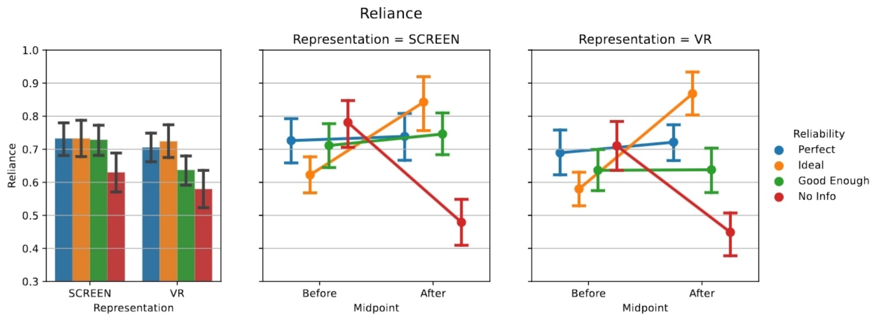

Figure 2: Amount of reliance. The representation factor is deconstructed.

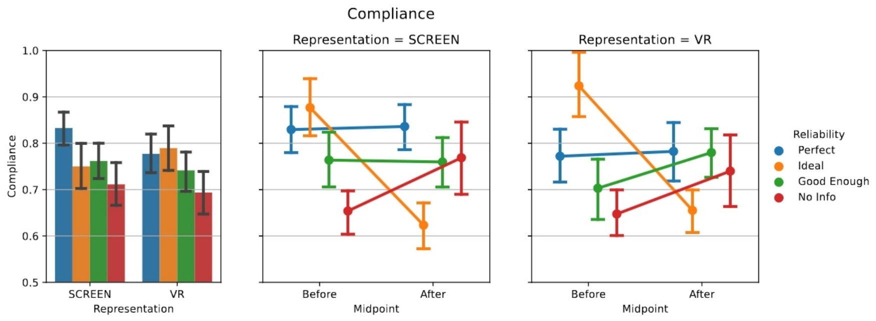

Figure 3: Amount of compliance. The representation factor is deconstructed.

Regarding the amount of deferrals, the ANOVA revealed only a main effect in the Time factor (F(1, 25) = 18.04, p < 0.001), along with a near-significant interaction with the Representation factor (F(1, 25) = 3.43, p = 0.07). It is interesting to note that instances of deferral were higher when robots are embodied in the VR condition within Representation. Additionally, this interacts with Time, as the number of deferrals increased after the mid-block feedback, likely an effort from participants to increase their decision-making speed and rewards. If we plot the amount of deferred trials over time, we can see a trending increase that is higher in the VR Representation condition.

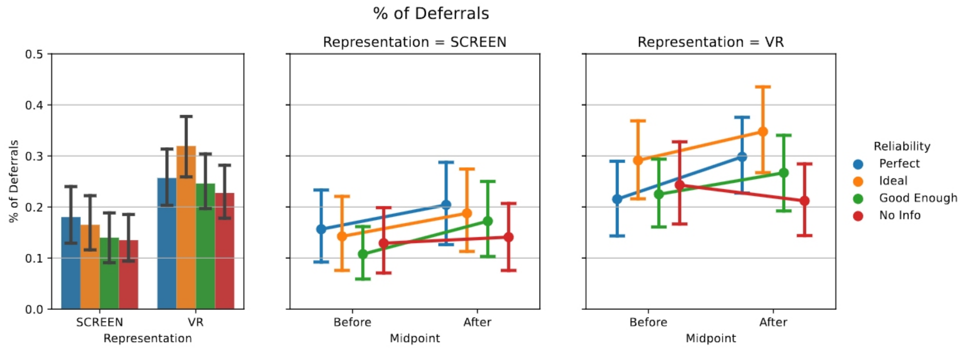

Figure 4: Proportion of deferred orders. The representation factor is deconstructed. The VR condition has a higher number of deferred trials.

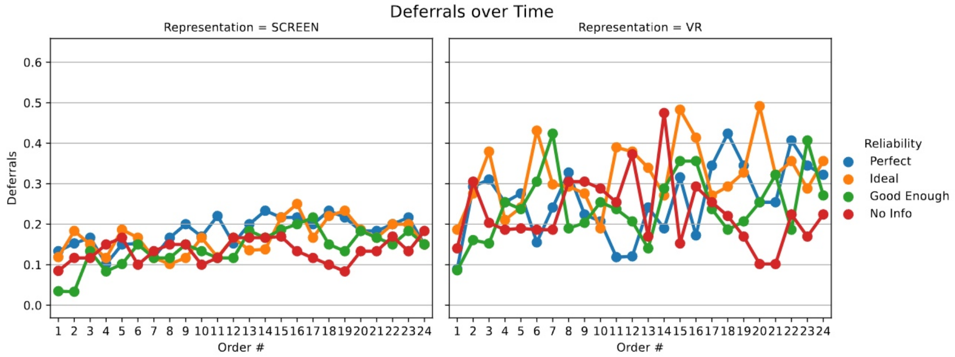

Figure 5: Deferred trials across time. The VR condition had a greater variance in deferred trials.

CONCLUSION

We conducted a study to investigate the effect of reliability and embodiment on performance and trust (n = 119), to fill the research gap on how dynamics in the human-agent teaming paradigm holds within a human-robot collaboration context, which is one of the representative goals in the development of AI agents. We observe that at 91% reliability, reliance drastically increases after the mid-block feedback (compliance changing symmetrically opposite). We infer that after the mid-block feedback, participants began to send more packages in an attempt to complete more orders, as they deemed the reliability satisfactory to deal with some errors. This behavior remained constant across representations, indicating what decisions are made are independent of embodiment, but tied closer to reliability. Closer tied to embodiment is how we make decisions: the results show that the number of deferrals increased steadily over time with embodied agents. We expected deferrals to increase as participants completed orders, yet a statistically significant increase in the VR Representation condition was observed. The combination of these findings reveals 2 major insights: 1) when an agent is presented as imperfect, embodiment changes trust calibration, as users defer their decisions to the agent to a greater degree, regardless of reliability. Embodiment could then allow for a higher level of trust calibration, resulting in more accurate and appropriate deferrals; and 2) humans can reasonably calibrate their expectations over AI systems with little to no feedback or transparency, with a single point of feedback being enough to cause significant changes in behavior. Future work should focus on comparing the results from a similar decision-making task (such as the one presented in this study) with a physical robot, to validate how decision-making is processed in the real world.

REFERENCES

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *34*(2), 181–186. https://doi.org/10.1109/TSMCC.2004.826268

Chien, S. Y., Lewis, M., Sycara, K., Liu, J. S., & Kumru, A. (2018). The effect of culture on trust in automation: Reliability and workload. *ACM Transactions on Interactive Intelligent Systems*, *8*(4). https://doi.org/10.1145/3230736

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Dixon, S. R., & Wickens, C. D. (2006). Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822

Esterwood, C., Essenmacher, K., & Yang, H. (2021). A meta-analysis of human personality and robot acceptance in human-robot interaction. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445542

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge, W., & Williams, M. A. (2018). Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System. *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, *August*, 7–14. https://doi.org/10.1109/ROMAN.2018.8525581

Hertzum, M., Andersen, H. H. K., Andersen, V., & Hansen, C. B. (2002). Trust in information sources: Seeking information from people, documents, and virtual agents. *Interacting with Computers*, *14*(5), 575–599. https://doi.org/10.1016/S0953-5438(02)00023-1

Kulms, P., & Kopp, S. (2016). The effect of embodiment and competence on trust and cooperation in human–agent interaction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10011 LNAI*, 75–84. https://doi.org/10.1007/978-3-319-47665-0_7

Lawson-Guidigbe, C., Louveton, N., Amokrane-Ferka, K., LeBlanc, B., & Andre, J. M. (2020). Impact of Visual Embodiment on Trust for a Self-driving Car Virtual Agent: A Survey Study and Design Recommendations. *Communications in Computer and Information Science*, *1226 CCIS*, 382–389. https://doi.org/10.1007/978-3-030-50732-9_51

Lee-Cultura, S., & Giannakos, M. (2020). Embodied Interaction and Spatial Skills: A Systematic Review of Empirical Studies. *Interacting with Computers*, *32*(4), 331–366. https://doi.org/10.1093/iwcomp/iwaa023

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Mollahosseini, A., Abdollahi, H., Sweeny, T. D., Cole, R., & Mahoor, M. H. (2018). Role of embodiment and presence in human perception of robots' facial cues. *International Journal of Human-Computer Studies*, *116*, 25–39. https://doi.org/10.1016/j.ijhcs.2018.04.005

Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. *ACM/IEEE International Conference on Human-Robot Interaction*, 33–42. https://doi.org/10.1145/3319502.3374839

Podevijn, G., O'Grady, R., Mathews, N., Gilles, A., Fantini-Hauwel, C., & Dorigo, M. (2016). Investigating the effect of increasing robot group sizes on the human psychophysiological state in the context of human–swarm interaction. *Swarm Intelligence*, *10*(3), 193–210. https://doi.org/10.1007/s11721-016-0124-3

Rae, I., & Takayama, L. (2013). *In-body Experiences: Embodiment, Control, and Trust in Robot-Mediated Communication*. 1921–1930.

Rice, S., Hughes, J., McCarley, J. S., & Keller, D. (2008). Automation Dependency and Performance Gains under Time Pressure. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1326–1329. https://doi.org/10.1177/154193120805201905

Saffo, D., Yildirim, C., Di Bartolomeo, S., & Dunne, C. (2020, April 25). Crowdsourcing virtual reality experiments using VRChat. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3334480.3382829

Schaffer, J., Humann, J., O'Donovan, J., & Höllerer, T. (2020). Quantitative Modeling of Dynamic Human-Agent Cognition. *Contemporary Research*, 137–186. https://doi.org/10.1201/9780429459733-7

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., & Young, J. E. (2015). Poor Thing! Would You Feel Sorry for a Simulated Robot?: A comparison of empathy toward a physical and a simulated robot. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 125–132. https://doi.org/10.1145/2696454.2696471

Tian, L., Carreno-Medrano, P., Allen, A., Sumartojo, S., Mintrom, M., Coronado Zuniga, E., Venture, G., Croft, E., & Kulic, D. (2021). Redesigning Human-Robot Interaction in Response to Robot Failures: A Participatory Design Methodology.

Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/3411763.3443440

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., & Tielman, M. L. (2020). Taxonomy of trust-relevant failures and mitigation strategies. *ACM/IEEE International Conference on Human-Robot Interaction*, 3–12. https://doi.org/10.1145/3319502.3374793

Tonkin, M., Vitale, J., Ojha, S., Clark, J., Pfeiffer, S., Judge, W., Wang, X., & Williams, M.-A. (2017). *Embodiment, Privacy and Social Robots: May I Remember You?* (pp. 506–515). https://doi.org/10.1007/978-3-319-70022-9_50

van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., & Haselager, P. (2014). Do Robot Performance and Behavioral Style affect Human Trust?: A Multi-Method Approach. *International Journal of Social Robotics*, 6(4), 519–531. https://doi.org/10.1007/s12369-014-0231-5

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Matarić, M. J. (2006). The role of physical embodiment in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 117–122. https://doi.org/10.1109/ROMAN.2006.314404

Walters, M. L., Koay, K. L., Syrdal, D. S., Dautenhahn, K., & Te Boekhorst, R. (2009). Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials. *Adaptive and Emergent Behaviour and Complex Systems - Proceedings of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2009*, 136–143.

Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018). Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams. In *Lecture Notes in Computer Science* (Vol. 10809, pp. 56–69). https://doi.org/10.1007/978-3-319-78978-1_5

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852

Ziemke, T. (2001). *Disentangling Notions of Embodiment*.