Does human-robot trust need reciprocity?

Joshua Zonca [1], Alessandra Sciutti [1]

1. Italian Institute of Technology, Cognitive Architecture for Collaborative Technologies (CONTACT) Unit

Summary

Trust is crucial in human-human and human-robot interactions. Among humans, trust is a relational phenomenon that requires reciprocity. However, research in human-robot interaction (HRI) has mostly relied on a unidirectional view of trust that focuses on robots' performance and reliability. The current work argues that reciprocity may also play an important role in the emergence of mutual trust and successful collaboration between humans and robots. We will gather works that reveal a reciprocal dimension in human-robot trust, discussing the implications for the development of "reciprocal" robots in HRI.

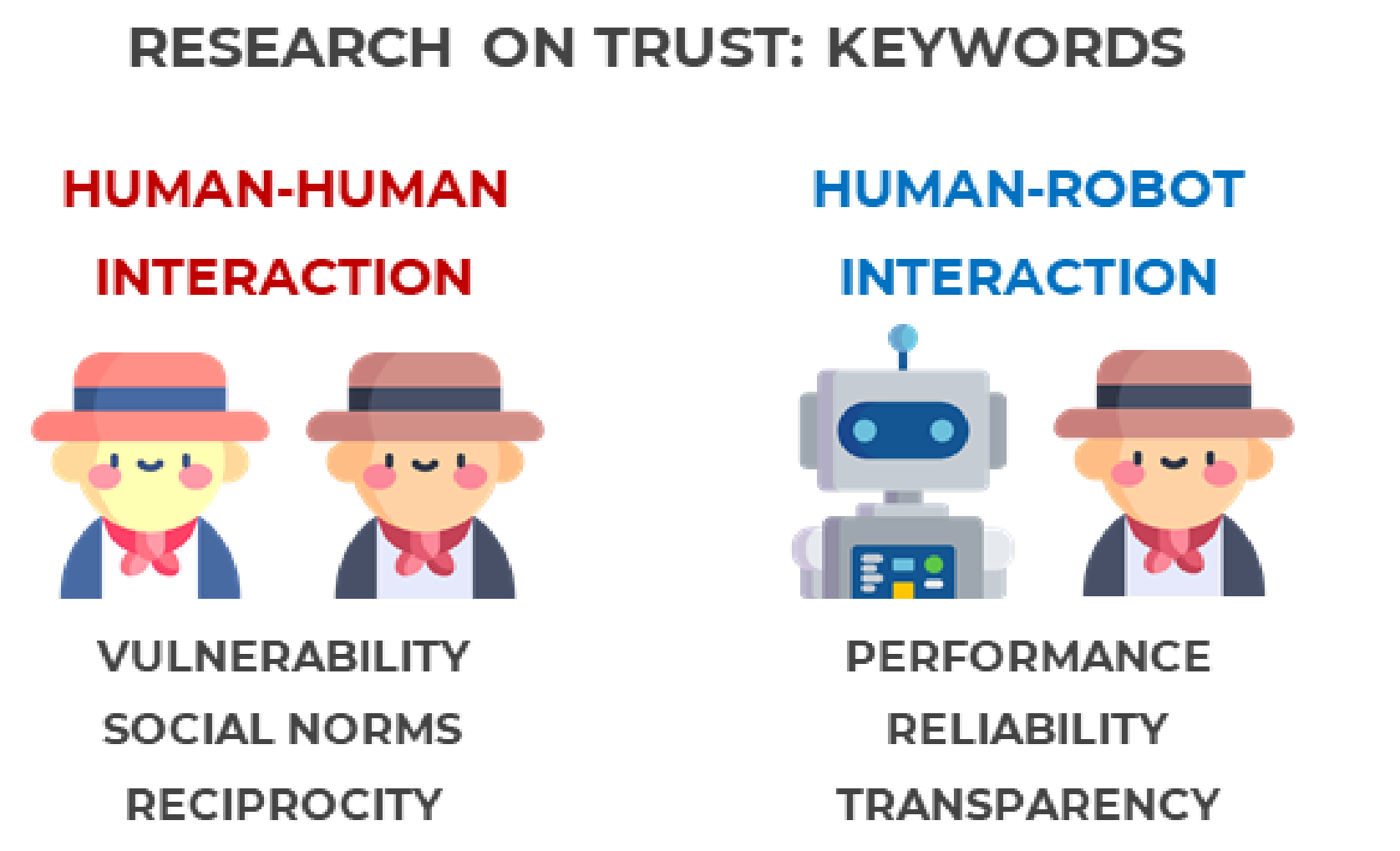
Introduction

A crucial mechanism sustaining the emergence and maintenance of cooperation among humans is reciprocity. Reciprocity is also fundamental for the maintenance of mutual trust between peers: if we do not trust others, it is unlikely that others will trust us in the future.

- Trust among humans is a relational and normative phenomenon [1]: all the individuals involved in interactions and relationships accept a condition of vulnerability to others.

However, reciprocity has not been given a central role in HRI research, especially in the study of human-robot trust.

- Trust in HRI is a unidirectional concept, which focuses on the physical, behavioral and functional characteristics of robots: humans trust them if they are functionally reliable and show high performance, whereas they do not trust them otherwise.


RESEARCH ON TRUST: KEYWORDS

HUMAN-HUMAN INTERACTION
VULNERABILITY
SOCIAL NORMS
RECIPROCITY

HUMAN-ROBOT INTERACTION
PERFORMANCE
RELIABILITY
TRANSPARENCY

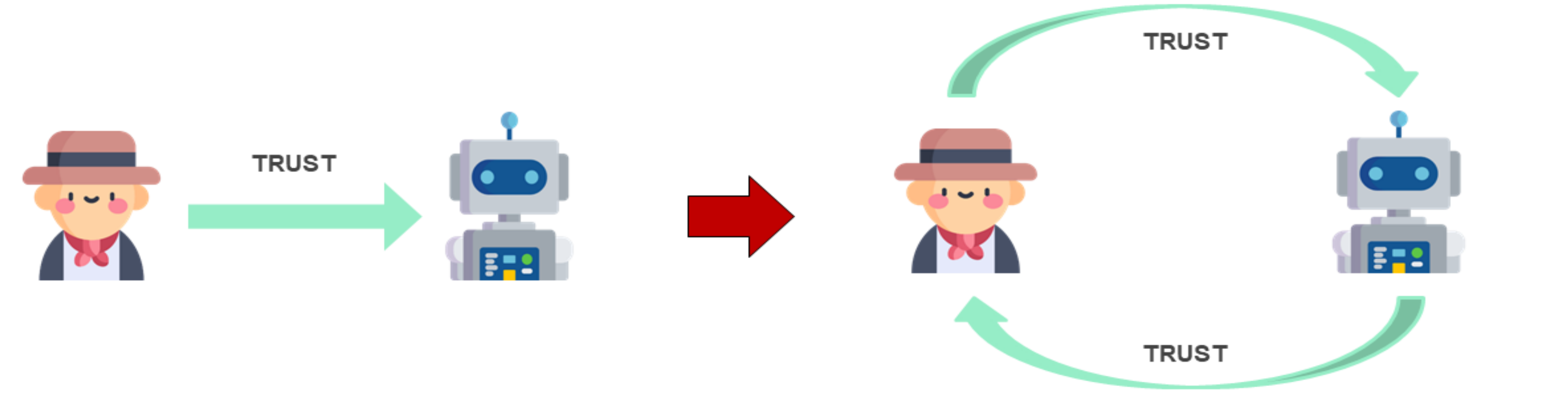
Research on trust in psychology/cognitive science and HRI have operationalized trust in different ways. The former studies trust as a relational phenomenon requiring reciprocity; the latter tends to interpret trust as a one-sided process of evaluation of robots' abilities: the more reliable robots are, the more they can be trusted.

Trust and reciprocity in HRI

Trust is one of the main mechanisms supporting collaboration with robots. Historically, research on trust in HRI conceptualized trust as a one-sided process of evaluation of the functional competence, performance and reliability of robotic agents (e.g. [2]). Nonetheless, recent evidence in HRI highlighted the emergence of distortions in the process of weighting of robots' competence and the relative expression of trust in them. Indeed:

- Trust towards robots is not always correlated with their reliability and performance [3-5].
- Humans show pro-social behavior towards robots (e.g., [6, 7]).
- Humans reciprocate with robots in repeated and multi-stage games such as the Prisoner's Dilemma and the Ultimatum Game [8].

Altogether, we hypothesize that trust-based relationships between humans and robots could be shaped, at least in part, by those relational and reciprocal mechanisms typically intervening in human-human interaction.

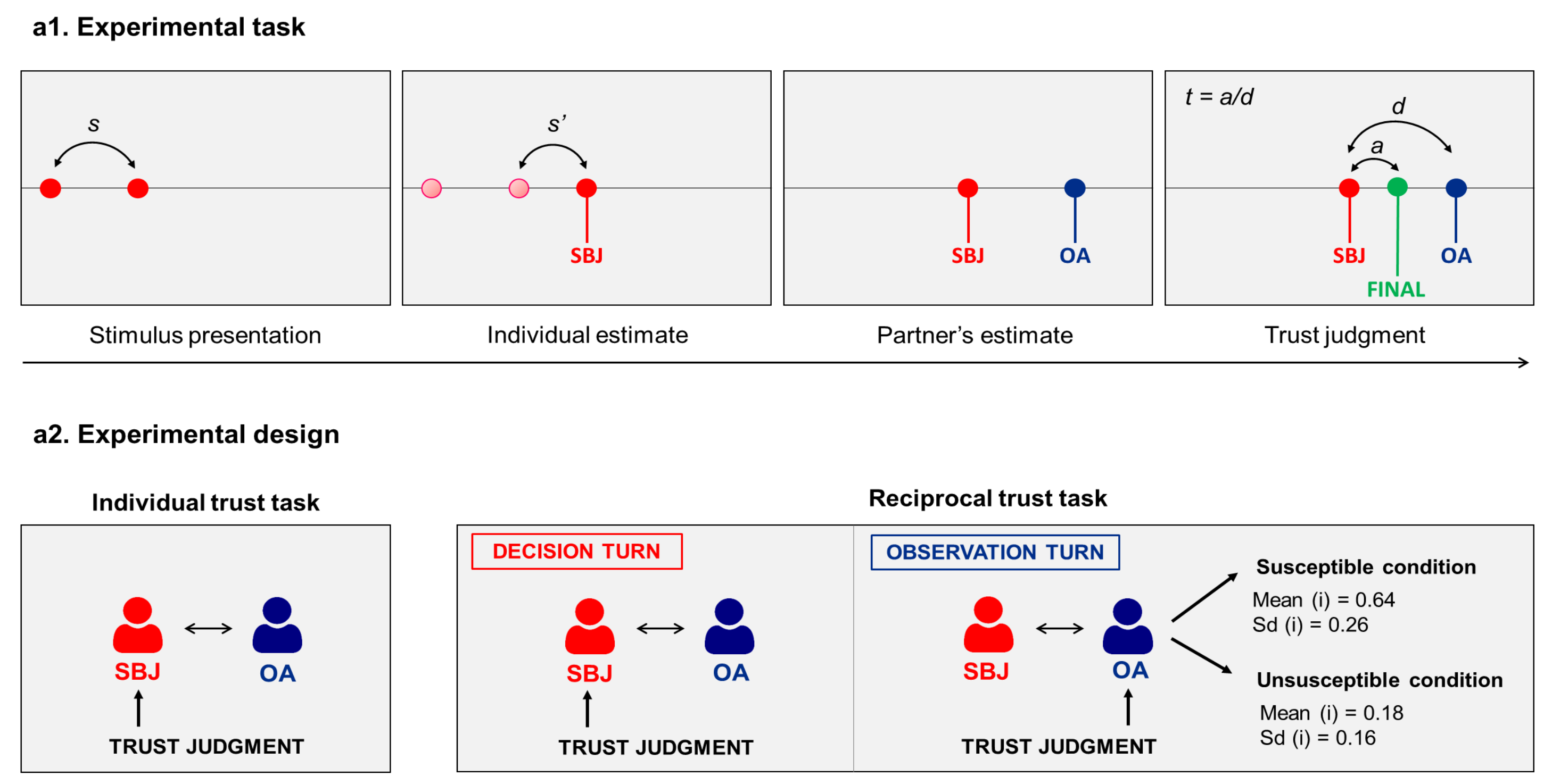


In fact, recent research in HRI put the accent on a bidirectional view of human-robot trust: our trust towards a robot may be influenced by the trust shown by the robot itself, following human-like reciprocal mechanisms:

1) A Robot showing vulnerability to humans or blaming itself for collaborative failures enhances human trust [9, 10].

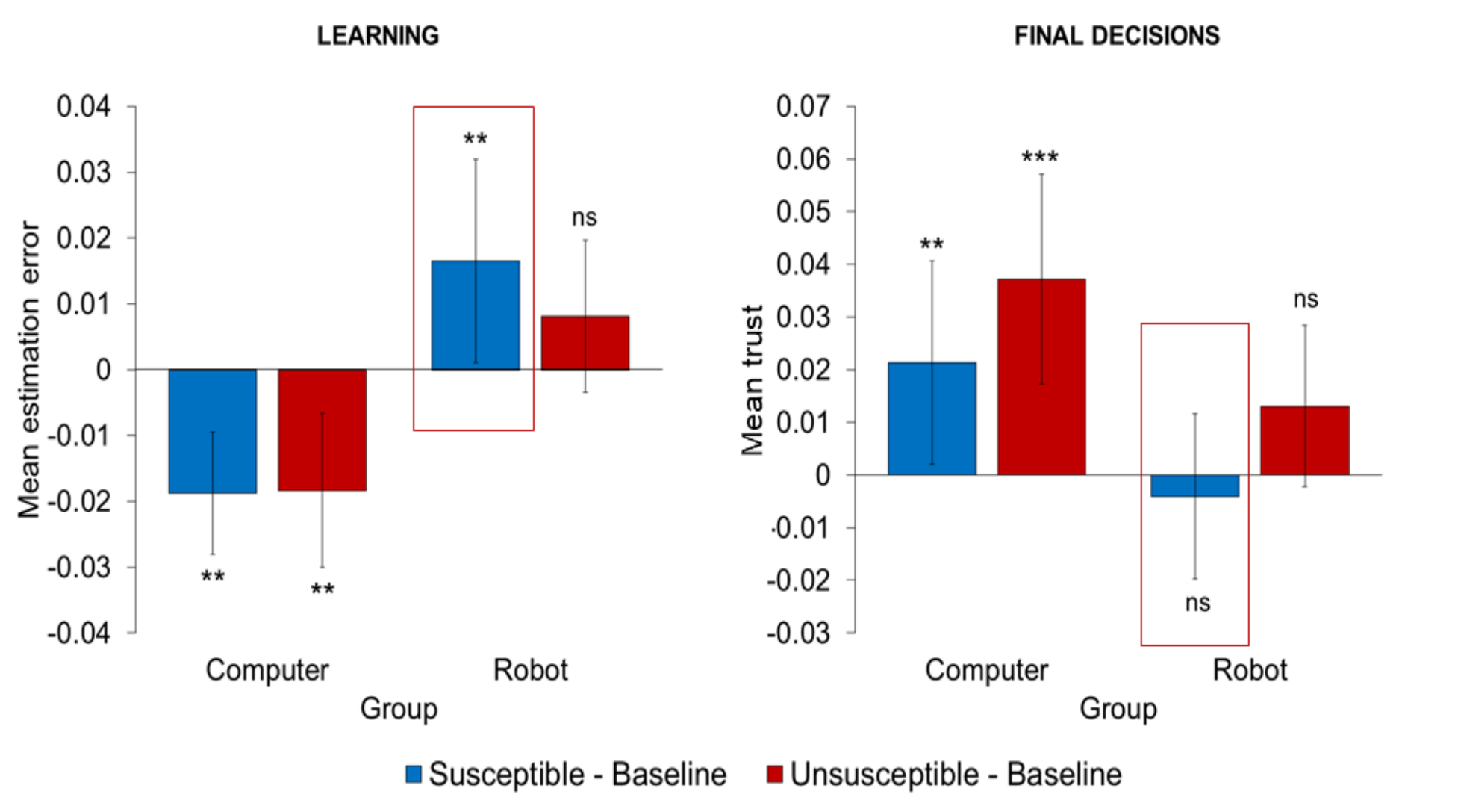
2) Cognitive architectures in which robots adapt their trust-related behavior to increase their trustworthiness [11].

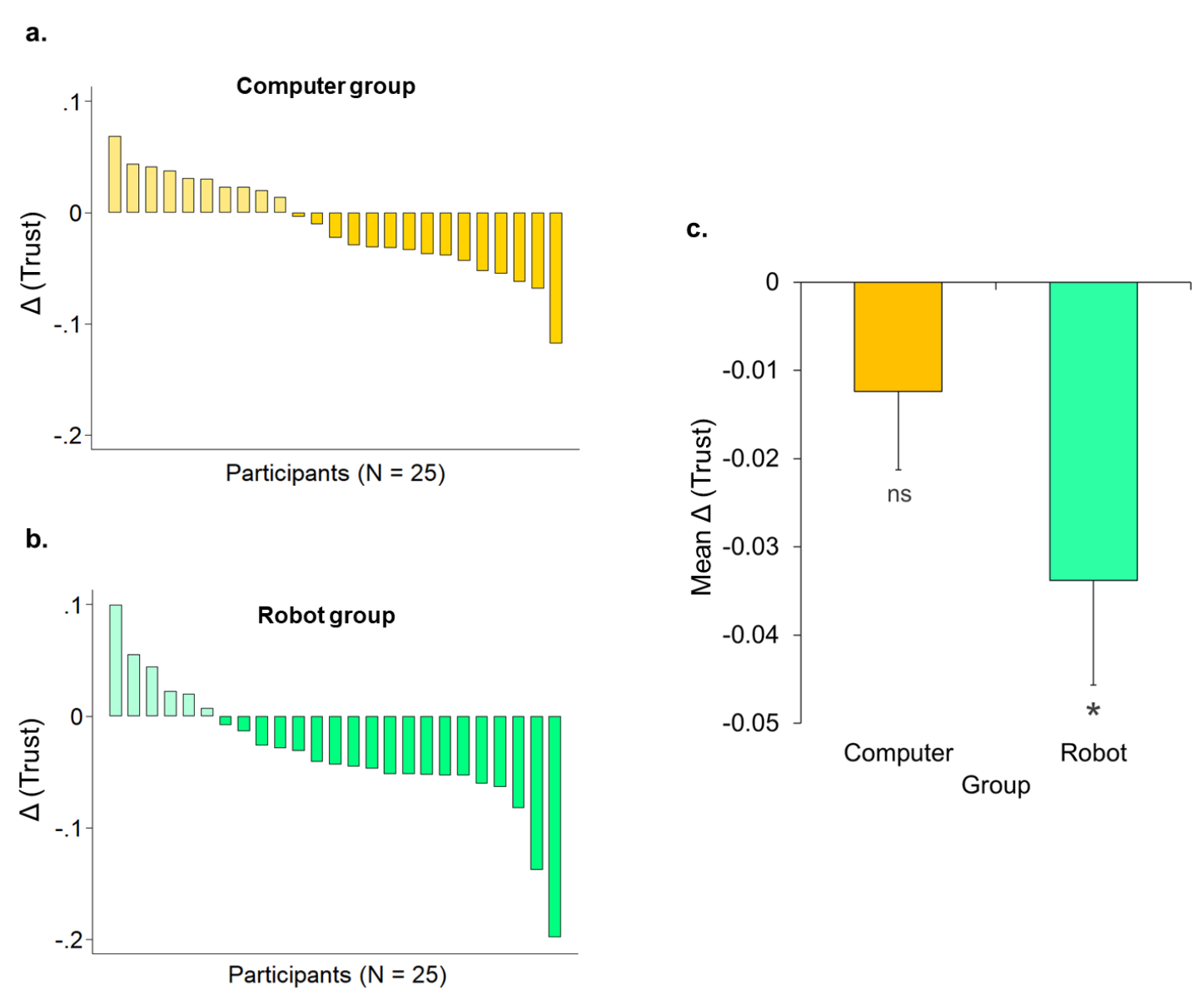
3) Investigating reciprocal trust in HRI: new results from our research group [12]



In a joint task (Reciprocal trust task), a participant and a humanoid robot iCub made perceptual judgments and signaled their trust in the partner. The robot's trust was manipulated along the experiment (Susceptible/Unsusceptible cond.) and participants could observe both robots' perceptual estimates and trust feedbacks (*t*). In a control condition (Individual trust task), only participants could reveal their trust in the partner. In a control group, participants performed the task with a computer.

Results





Participants did not learn from a robot that was showing high trust in them, since the robot signaled incompetence. However, they were unwilling to disclose their distrust to the robot (red rectangles). *** p < 0.001, ** p < 0.01, * p < 0.05, ns: not significant, mixed-effects model.

If participants did not expect future interactions with the robot, their trust in the robot decreased (this did not happen with a computer). * p < 0.05, ns: not significant, Wilcoxon s-r test.

Implications for HRI and future directions

In order to design and develop "reciprocal" robots:

- A social and collaborative robot should be able to adapt its trust-related behavior to modulate and maximize human trust. It should decide when to take the lead and when to comply with the human partner during a joint task, in order to optimize task-related performance and preserve human-robot trust-based collaboration and social norms.
- Robots should be able to behave as adaptive agents involved in relationally-rich interactions, with their high-level goals and motives.

These considerations open several questions for HRI:

- What should be the robots' goals in interaction? And should humans have specific knowledge about them?
- How should robots balance conflicting goals taking into consideration relational and functional considerations?
- Shall we design robots that can exhibit negative reciprocity towards humans?

Conclusions

The ambition of designing robotic collaborators, rather than anthropomorphic mechanical tools, opens the question of whether human-robot trust relationships should be reciprocal, as those among human peers. Although research on the role of reciprocity in human-robot trust is very limited, recent findings suggest that trust towards robots is not a mere function of their perceived competence and reliability. Further research is needed to unveil the extent to which human trust in robots can be shaped by relational and reciprocal dynamics in joint tasks. These aspects may be fundamental in the design of robots that act as collaborators in contexts such as healthcare, rehabilitation, education and assistance for the elderly.

References

[1] M. A. Nowak, "Five rules for the evolution of cooperation," Science, vol. 314, pp. 1560-1563, 2006.
[2] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," Hum. Factors, vol. 53, pp. 517-527, 2011.
[3] M. Salem, G. Lakatos, F. Amirabdollahian and K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 1-8, 2015.
[4] A. M. Aroyo, F. Rea, G. Sandini and A. Sciutti, "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?," IEEE Robot. Autom. Lett., vol. 3, pp. 3701-3708, 2018.
[5] P. Robinette, W. Li, R. Allen, A. M. Howard and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 101-108, 2016.
[6] J. Connolly, V. Mocz, N. Salomons, J. Valdez, N. Tsoi, B. Scassellati and M. Vázquez, "Prompting prosocial human interventions in response to robot mistreatment," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 211–220, 2020.
[7] R. Oliveira, P. Arriaga, F. P. Santos, S. Mascarenhas and A. Paiva, "Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behavior," Comput. Hum. Behav., 106547, 2021.
[8] E. B. Sandoval, J. Brandstetter, M. Obaid and C. Bartneck, "Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game," Int. J. Soc. Robot., vol. 8, pp. 303-317, 2016.
[9] S. Strohkorb Sebo, M. Traeger, M. Jung and B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 178-186, 2018.
[10] D. P. Van der Hoorn, A. Neerincx and M. M. de Graaf, "I think you are doing a bad job! The Effect of Blame Attribution by a Robot in Human-Robot Collaboration," ACM/IEEE Int. Conf. Human-Robot Interact., pp. 140-148, 2021.
robots and virtual agents to trigger prosocial behavior," Comput. Hum. Behav., 106547, 2021.
[11] S. Vinanzi, A. Cangelosi and C. Goerick, "The collaborative mind: intention reading and trust in human-robot interaction," iScience, vol. 24, 102130, 2021.
[12] J. Zonca, A. Folsø and A. Sciutti, "If you trust me, I will trust you: the role of reciprocity in human-robot trust," submitted for publication, arXiv preprint: http://arxiv.org/abs/2106.14832

Acknowledgments

The authors gratefully acknowledge the financial support of the European Research Council (ERC Starting Grant 804388, wHiSPER).

Contact

Joshua Zonca
Email: joshua.zonca@iit.it