Nobody's Perfect

On Trust in Social Robot Failures

Krantz, Amandus; Haresamudram, Kashyap; Balkenius, Christian

Nobody's perfect:
On Trust in Social Robot Failures

Kashyap Haresamudram[*]
kashyap.haresamudram@lth.lu.se
Lund University
Lund, Sweden

Amandus Krantz[*]
amandus.krantz@lucs.lu.se
Lund University
Lund, Sweden

Christian Balkenius
christian.balkenius@lucs.lu.se
Lund University
Lund, Sweden

ABSTRACT

With robots increasingly succeeding in exhibiting more human-like behaviours, humans may be more likely to 'forgive' their errors and continue to trust them as a result of ascribing higher, more human-like intelligence to them. If an integral aspect of successful HRI is to accurately communicate the competence of a robot, it can be argued that the technical success of the robot in exhibiting human-like behaviour can, in some cases, lead to a failure of the interaction by resulting in misperceived human-like competence. We highlight this through the example of speech in robots, and discuss the implications of failures and their role in HRI design.

CCS CONCEPTS

• **Computing methodologies → Cognitive science**; • **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

robotics, trust, human-robot interaction

ACM Reference Format:
Kashyap Haresamudram, Amandus Krantz, and Christian Balkenius. 2023. Nobody's perfect: On Trust in Social Robot Failures. In *Proceedings of The Imperfectly Relatable Robot.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The general conception of a robot in popular culture often tends to land close to what roboticists would now identify as a humanoid, social robot. While these types of robots were relegated to imaginings and fictions for a long time, in recent years we have seen a proliferation of research into robots that look and act like humans (though they are far from replicating the full complexity of human behaviour). This development has brought forth an urgency in understanding what interaction between such agents and humans might look like. It has never before been possible or important to understand how humans would perceive other intelligent beings that embody human qualities, and how this perception would shape interactions with them.

Recently, one fundamental socio-cognitive phenomenon that is receiving a lot of attention, not only in robotics but technology in general, is trust. In many ways, human society is built around trust, and human interaction is profoundly shaped by it [7]. As a crucial social phenomenon, trust is a central concept in social robotics (irrespective of whether they are humanoid robots). Understanding how social interaction shapes human perception and informs trust, and how this relationship translates to human-robot interaction (HRI) is essential to the development of social robots.

Currently, there are no robots that can fully reproduce the complexity of human social behaviour; the ones that do exhibit what could reasonably be termed 'social' behaviour are highly limited or specialized to perform certain tasks. Despite the limitations, they can often fail at performing the tasks consistently. It then becomes important to understand how robot failures impact the perception of sociability (or competence) of the robot, and how this influences perceptions of trust and trustworthiness. In this abstract we focus on one aspect of this dynamic by presenting our thoughts on robots with speech being 'forgiven' for failures, how this affects trust and reliance in failing robots, and the role of less-than-perfect robot behaviour in HRI.

2 PERCEIVED-INTELLIGENCE AND TRUST

Complex speech is an ability that, at least in humanly comprehensible forms, is only found in humans and might thus be considered a sort of signal of "human-like intelligence". In [6], we showed that possessing the ability to speak can almost completely remove any negative effects that may be caused by a robot's failure in operation and that speaking robots have an advantage over non-speaking ones, in that they are able to maintain their perceived trustworthiness despite both failing at given tasks. Similar effects have also been seen in other studies (e.g. [1, 10]) where robots that were able to apologize for failing and explain why the failure happened were also able to reduce loss of trust. We argued then that this effect could potentially be linked to speech increasing the perceived intelligence of the robot, since perceived intelligence/competence has been correlated with trust in the past [3].

Theoretically, though, trust is linked to reliance. There are several perspectives on the relationship between trust and reliance in human interaction [11]. Doxastic accounts of trust hold that, from the point of view of a truster, trust is a species of reliance involving belief that the trustee will do as they are relied upon to do. Hardin [5] elaborates on this notion through his 'Encapsulated Interest' account of trust, where the truster's interests are encapsulated within the trustee's incentives to do as they are relied upon to do [4]. Considering the above, an unreliable robot by way of failing at tasks, whether it is able to speak, should exhibit a consistent level of

[*] Both authors contributed equally to this research.

The Imperfectly Relatable Robot, March 13, 2023, Stockholm, Sweden

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

loss in trust with its non-speaking counterpart. The contradictory experimental observation of sustained trust in failing robots that speak poses a challenging question.

Implicit within several trust definitions is the notion that it is a social phenomenon that occurs between humans or humans and agents with some form of intelligence and autonomy (this is reflected in the aforementioned 'encapsulated trust' that accounts for the trustee's incentives). While alternate thoughts do exist, such as Castelfranchi and Falcone [2] who argue that trust even applies to objects such as instruments and technologies, in the case of robots that are relatively intelligent and can possess limited autonomy, conventional arguments about social trust still apply. Given this theoretical perspective on trust, if a robot's ability to speak is perceived as a sign of intelligence or autonomy, there is an argument to be made that it would lend itself to trust phenomenon as observed in human-human interaction, as well as accompanying phenomenon such as reliance.

We argue that an alternate explanation for sustained trust in failing robots with speech (as opposed to ones without speech) could also be that speech changes the *nature* of intelligence as perceived by humans. By way of exhibiting human-like abilities, such as speech, the robot may be perceived to possess a more human-like intelligence rather than some form of machine intelligence that is distinct (which is the case in reality). This perception of human-like intelligence may trigger social processes seen in human-human interaction, such as trust and reliance.

3 TRUST AND FAILURE IN HRI

Whether we should accept failure in robots is a topic that becomes more and more important as robots become more common in society. On the one hand, one might argue that the environments these robots are entering are too non-deterministic for it to be possible to foresee every possible outcome and implement recovery strategies. On the other hand, robots are engineered beings, and it might thus not be reasonable to hold them to the same standard as humans in their ability to err. Regardless of which side one takes in this discussion, it has to be acknowledged that the situation is currently closer to the former than the latter; errors and failures can, and do, happen in HRI today. This makes it important to understand how failure impacts a robot's perceived trustworthiness, and how different characteristics might interact with this impact.

While the exact nature of the relationship between trust and reliance is still up for debate, it is generally agreed that trust and reliance are positively correlated. It is then surprising to find that perceived trustworthiness of speaking robots is not affected by their failure at performing a given task. If one cannot rely on a robot to perform a task, then it should, in theory, reduce trust. This relationship between reliance and trust does hold true for mute robots, but not for speaking robots [6]. One explanation could be that because humans are not used to other beings exhibiting human-like qualities, the ability of speaking robots to do so inflates their perceived competence. As a result, we argue that, inadvertently, evaluation of the robot's competence is not only done on its ability to perform a given task but exhibit human qualities. Despite failing at their task, observers may still evaluate the competence of the robot to be high simply by way of their ability to exhibit human-like intelligence through speaking. Consequently, the robot retains its perceived trustworthiness. Additionally, humans are generally not seen as perfect beings and failures are not surprising, while mainstream depictions of robots often cause machine intelligence to be perceived as infallible, sometimes to a fault. With such a perception, seeing an agent with machine intelligence fail at a task would seem out of the ordinary. On the other hand, a robot perceived to possess human-like intelligence would instead be expected to fail sometimes in keeping with human expectations. Such an agent failing or performing less than perfectly would thus not be surprising (to a certain extent) and evaluations of its performance and trustworthiness would be more lenient than if it was perceived as having machine intelligence.

4 NUANCES IN HRI FAILURE

While robots that are malfunctioning should be repaired or discarded to prevent damage or harm, there is a case to be made for not completely removing trivial errors and failures from the operation of robots. For example, it has been shown that, while failures reduce performance in collaborative tasks, they also significantly increase positive emotions [9]. This has been linked to the *Pratfall Effect* [8] in which people with a high perceived competence become more likeable when they make mistakes, than if they were always performing perfectly. If the Pratfall Effect does apply to robots, there may be some benefit to not striving for perfect behaviour every single time, at least for robots which are intended to be more social than functional.

Emerging from this perspective is the notion that there are different types of failures for robots, and some can be beneficial. Failures that are attributed to competence and reliability generally erode trust, while failures that do not affect reliability of the robot to perform a certain task benefit the interaction. They add characteristics to robots that can be charming or perceived as an expression of personality, further adding to their human quality. Seeing as exhibiting human-like behaviours makes robots retain trust despite failing at tasks, this could be a useful buffer to set expectations about robot competence.

5 OPPORTUNITIES IN ROBOT FAILURE

Strategic use of less-than-perfect operation could be beneficial for the calibration of a robot's perceived capability with its actual capability. When a robot is perfect in some aspect, it is often assumed that it is also perfect in others. This does not always hold true, however, a bipedal robot that is good at locomotion in a straight line is not necessarily as good at, for example, turning. This could cause the user to put too much trust in the robot, causing more intense negative emotions when a failure does happen. As an extension, over-reliance poses a significant challenge in human-robot interaction. If humans attribute human-like intelligence to robots and inflate their competence as a consequence of their ability to display some human-like qualities, there is a risk of unwarranted, perceived trustworthiness.

However, there is an opportunity here to explore failure as a design tool to manage perceptions about robot competence and trustworthiness. Much research is needed to establish the limitations of failures, as well as expand upon the perspectives detailed above on the nuances of failures. Research is also needed on whether humans-like intelligence does in fact enable human interaction phenomenon to apply to HRI, and where the limitations lie. Lastly, the position presented here is primarily based on speech in robots, whether the same phenomenon can be reproduced with other uniquely human behaviours and attributes is unknown, presenting yet another avenue of further research.

6 CONCLUSION

In this abstract, we briefly introduced a study on perceived trustworthiness in speaking vs mute robots and highlighted inconsistencies in the findings of the study with theories on trust. We explained that if reliance and trust are related, it is odd that speech makes humans forgive errors made by robots. And we provided a possible theoretical explanation for the results by hypothesising that speech makes humans attribute human-like intelligence to robots, and this extends trust despite failures, as is normal to do with other humans within reason. We then briefly reflected on what attributing human-like intelligence would mean for failures in HRI, and highlighted potential areas of research to further investigate our position.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program - Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

REFERENCES

[1] David Cameron, Stevienna de Saille, Emily C. Collins, Jonathan M. Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh, and James Law. 2021. The Effect of Social-Cognitive Recovery Strategies on Likability, Capability and Trust in Social Robots. *Computers in human behavior* 114 (Jan. 2021), 11. https://doi.org/10.1016/j.chb.2020.106561

[2] Cristiano Castelfranchi and Rino Falcone. 2020. Trust: Perspectives in Cognitive Science. In *The Routledge Handbook of Trust and Philosophy*. Routledge.

[3] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of management annals* 14, 2 (March 2020), 627–660. https://doi.org/10.5465/annals.2018.0057

[4] Sanford C. Goldberg. 2020. Trust and Reliance 1. In *The Routledge Handbook of Trust and Philosophy*. Routledge.

[5] Russell Hardin. 1992. The Street-Level Epistemology of Trust. *Analyse & Kritik* 14, 2 (Nov. 1992), 152–176. https://doi.org/10.1515/auk-1992-0204

[6] Amandus Krantz, Christian Balkenius, and Birger Johansson. 2022. Using Speech to Reduce Loss of Trust in Humanoid Social Robots. In *SCRITA Workshop Proceedings (arXiv:2208.11090)*. IEEE, Naples, Italy, 4. arXiv:2208.13688 [cs]

[7] Stephen P. Marsh. 1994. *Formalising Trust as a Computational Concept.* Ph. D. Dissertation. University of Sterling.

[8] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI* 4 (2017), 15.

[9] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O. Arras. 2016. Errare Humanum Est: Erroneous Robots in Human-Robot Interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 501–506. https://doi.org/10.1109/ROMAN.2016.7745164

[10] Alessandra Rossi, Fernando Garcia, Arturo Cruz Maya, Kerstin Dautenhahn, Kheng Lee Koay, Michael L. Walters, and Amit K. Pandey. 2019. Investigating the Effects of Social Interactive Behaviours of a Robot on People's Trust during a Navigation Task. In *Towards Autonomous Robotic Systems (Lecture Notes in Computer Science)*, Kaspar Althoefer, Jelizaveta Konstantinova, and Ketao Zhang (Eds.). Springer International Publishing, Cham, 349–361. https://doi.org/10.1007/978-3-030-23807-0_29

[11] Judith Simon. 2020. *The Routledge Handbook of Trust and Philosophy.* Routledge.