# Evaluating Trust and Safety in HRI: Practical Issues and Ethical Challenges

Maha Salem
University of Hertfordshire
College Lane, Hatfield
AL10 9AB, United Kingdom
m.salem@herts.ac.uk

Kerstin Dautenhahn
University of Hertfordshire
College Lane, Hatfield
AL10 9AB, United Kingdom
k.dautenhahn@herts.ac.uk

## ABSTRACT

In an effort to increase the acceptance and persuasiveness of socially assistive robots in home and healthcare environments, HRI researchers attempt to identify factors that promote human trust and perceived safety with regard to robots. Especially in collaborative contexts in which humans are requested to accept information provided by the robot and follow its suggestions, trust plays a crucial role, as it is strongly linked to persuasiveness. As a result, human-robot trust can directly affect people's willingness to cooperate with the robot, while under- or overreliance could have severe or even dangerous consequences. Problematically, investigating trust and human perceptions of safety in HRI experiments is not a straightforward task and, in light of a number of ethical concerns and risks, proves quite challenging. This position statement highlights a few of these points based on experiences from HRI practice and raises a few important questions that HRI researchers should consider.

## Keywords

Socially Assistive Robots; Cooperation; Trust and Safety

## 1. INTRODUCTION & BACKGROUND

As robots are increasingly designed and developed to assist humans with everyday tasks in their homes and in healthcare settings, questions regarding the robot's safety and trustworthiness are inevitably to be addressed. In a possible future scenario, a home companion robot may remind an elderly person to take their medication or to get physically active by suggesting some exercise on a regular basis. Since such interactions, particularly in the domestic domain, are intended to take place in an informal and unstructured way and without any expert supervision, roboticists and human-robot interaction (HRI) researchers face multiple challenges.

Besides work on the robot's technical reliability and operational safety to establish and maintain effective relationships with assistive robots, another crucial factor is *trust* [3]. Not only does trust play an important role in human interactions (in particular with regard to critical decisions), but it could also increase the robot's acceptance in its role as a collaborative partner [4].

Moreover, since trust is strongly linked to persuasiveness in social and collaborative contexts, it could directly affect people's willingness to cooperate with the robot, e.g. by accepting information or following its suggestions [2]. As a result, robot designers have set out to develop machines that act socially in a way such that humans perceive them as safe and trustworthy. Problematically though, inappropriate levels of trust with regard to the robot could not only result in a frustrating HRI experience, but under- and overreliance could even bear serious consequences [3].

For example, a person doubting the robot's competence and thus not willing to rely on its recommendations might refuse to take their medication in time following the robot's reminder. On the other hand, a person overrelying on the robot might ignore signs of malfunction, e.g. in the form of a sensor failure, and put their own safety at risk when asking the robot to grasp and carry a hot beverage for them.

Despite its importance, investigating and successfully measuring trust and human perceptions of safety in HRI remains an extremely challenging task which bears a number of ethical concerns and risks. Specifically, how can HRI researchers design meaningful experimental scenarios that take place in natural environments and test realistic aspects of safety and trust, without potentially putting their participants at risk? This paper aims to stimulate discussion within the wider community by highlighting a few of the challenges and issues related to safety- and trust-related HRI research.

## 2. CHALLENGES IN HRI RESEARCH ON SAFETY AND TRUST

In an attempt to investigate human-robot trust, an experiment recently conducted by Salem et al. [7] as part of the EPSRC funded project on "Trustworthy Robotic Assistants" provided interesting insights regarding the complexities of the concept of trust in the social HRI context: not only do definitions of trust in the literature often lack agreement and generalization [1], but also its quantification by means of experimental measures proves extremely difficult and – depending on the variables used – sometimes contradictory.

The study built on previous work that identified reliability and predictability as two main promoting factors of trust: Muir and Moray [5], for example, propose that trust is mainly based on the extent to which the machine is perceived

*The Emerging Policy and Ethics of Human Robot Interaction,* March 2, 2015, Portland, Oregon, USA.

to perform its function properly, implying that machine errors can strongly affect trust. More specifically, according to Corritore et al. [1], an accumulation of small errors seems to have a more severe and long-lasting impact on trust than a single large error.

Inspired by these findings, Salem and colleagues designed their experimental study in a home environment and manipulated the robot's behavior in a correct vs. faulty condition, while tapping different dimensions of trust based on a variety of unusual collaborative tasks [7]. Trust was measured using self-reported quantitative and qualitative questionnaire data, as well as behavioral data assessing cooperation with the robot as a "behavioral outcome of trust" [8].

In summary, the study revealed that while subjective measures of self-reported questionnaire data showed a significant effect of condition, participants in both conditions did not differ objectively in their willingness to comply with the robot's unusual requests. That is, despite dealing with a clearly faulty robot, participants still followed the robot's instructions which – within the experimental scenario – would lead to damaged property and breaches of privacy.

Participants' interview data as well as feedback from the reviewers of the conference paper describing the study ([7]) highlighted some of the main challenges when conducting this type of research, which can be summarized as follows:

- *Ethical issues and legal boundaries.* One reviewer remarked that trust requires participants to have something at stake or that they perceive a certain risk in the situation. However, a truly 'risky' experimental scenario is very unlikely to receive ethics approval from the review board. Consequently, HRI researchers have very limited means of measuring trust (and importantly under- or overreliance) in experimental scenarios that bear a realistic safety hazard. Equally, it would be unethical to deceive participants by telling them that they are going to interact with an unsafe or faulty robot with limited controllability, as this could put them into an unwarrantably stressful situation.

- *Experimental observer and novelty effect.* Participants are aware of the fact that they are part of an experiment:
  - Several participants reported that they followed the robot's instructions "because it was an experiment".
  - Some participants admitted that they would have done anything the robot asked them to do (with a few people referring to themselves as having been in "autopilot mode"), as they were completely absorbed by the novelty of the experience.
  - Some might consider the robot to simply represent or be an extension of the researcher/programmer, i.e. perceiving it as a remote-controlled entity rather than an autonomous agent.
  - Even if the designed collaborative task did impose a realistic risk on participants, they might still feel "safe" as they know they are part of an approved study associated with an established university or lab.

These observations make clear that there are some critical limitations that hinder HRI researchers from establishing a realistic understanding of potential risks related to uncalibrated human-robot trust and perceived safety. At the same time, however, the study described above highlighted the participants' alarming willingness to blindly follow a (faulty) robot, and it remains unclear whether one could expect to find the trend of such an "autopilot mode" also in non-experimental or long-term interactions. For example, one study participant stated "you trust the robot has been programmed appropriately and accordingly to do the right thing. I would expect of a robot to always give me the right answer and the right thing."

Many people already commonly rely on GPS devices to guide them by providing directions while driving, with suboptimal routes, detours or even errors in the route-planning remaining undetected at best, or resulting in dangerous incidents at worst. Problematically, in a home care scenario such overreliance could, for instance, result in an elderly person with dementia taking an overdose of medication if a malfunctioning robot reminds the user of the same scheduled dose intake multiple times.

Therefore, and in view of the possibly serious consequences particularly with regard to vulnerable people, a clear understanding of the dynamics and potential risks involved in the development of trust in HRI is essential before socially assistive robots can be placed into people's homes.

## 3. LOOKING AHEAD AND BEYOND LAB RESEARCH

Riek and Howard [6] suggest that "situations in which ethical problems are noticed only after the fact" should be avoided. Therefore, and to complement the perspective based on the above described experimental findings and insights, our considerations should ideally go beyond lab-related research while still at the developmental stage. In the following, a (non-exhaustive) list of questions that should already be dealt with now is proposed.

- How much "safety" regarding home companion robots can robot designers and manufacturers really guarantee, especially if the robot is equipped with some level of autonomy and/or learning capability? In this context, would it be appropriate to differentiate between *safe hardware* vs. *safe software* vs. *safe interactions*?

- What can such robots and the risks they might bear be compared to in today's households? If we look at other devices currently "approved" for home use, how do they differ from our vision of robot companions in the house (e.g. not autonomous/mobile/multi-purpose/able to 'learn')?

- Even if it was possible to certify a homecare robot as safe, there may be a discrepancy between this *certified safety* and its *perceived safety*: that is, a certified robot might be considered safe objectively, but a (non-expert) user may still perceive it as unsafe or scary. What role does the robot's design play in this respect? And how likely are these initial perceptions going to change in long-term interactions (e.g. due to adaptation/habituation), especially when people experience how (un)safe the robot really is?

- Since the target group of companion robots are typically non-expert users who potentially belong to a vulnerable and dependent population, should the use of such robots require compulsory training/licenses, in a similar fashion as required to use cars or industrial robots?

- Inspired by current debates about the safety and ethical implications of self-driving cars, should we as researchers in this area also develop a vision of how "safe" such robots intended for use in unstructured and unsupervised home environments can realistically ever be? If so, how do these predictions compare to other areas of HRI in which potentially autonomous robots act in similarly complex settings in close proximity to humans (e.g. search and rescue)?

These and other questions should be discussed in the context of ethics and user safety to raise awareness and promote experimental guidelines within the HRI community, so that this line of research can advance *before* robots are actually deployed in the homes of vulnerable populations.

## 4. ACKNOWLEDGMENT

The authors were partially supported by EPSRC grant EP/K006509.

## 5. REFERENCES

[1] C. L. Corritore, B. Kracher, and S. Wiedenbeck. On-line trust: Concepts, evolving themes, a model. *Int. J. Hum.-Comput. Stud.*, 58(6):737–758, 2003.

[2] A. Freedy, E. de Visser, G. Weltman, and N. Coeyman. Measurement of trust in human-robot collaboration. In *International Symposium on Collaborative Technologies and Systems (CTS 2007)*, pages 106–114, 2007.

[3] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. de Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527, 2011.

[4] J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno. Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4(893), 2013.

[5] B. M. Muir and N. Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.

[6] L. D. Riek and D. Howard. A code of ethics for the human-robot interaction profession. In *Proceedings of We Robot 2014*, 2014.

[7] 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. In *10th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015)*, 2015.

[8] J. M. Wilson, S. G. Straus, and B. McEvily. All in due time: The development of trust in computer-mediated and face-to-face teams. *Organizational Behavior and Human Decision Processes*, 99(1):16–33, January 2006.