Towards Safe and Trustworthy Social Robots: Ethical Challenges and Practical Issues

Maha Salem[1], Gabriella Lakatos[2],
Farshid Amirabdollahian[3], Kerstin Dautenhahn[4]

Adaptive Systems Research Group, University of Hertfordshire, UK
[[1]me@mahasalem.net, [2]gabriella.lakatos@gmail.com,
[3]f.amirabdollahian2@herts.ac.uk, [4]k.dautenhahn@herts.ac.uk]

Abstract. As robots are increasingly developed to assist humans socially with everyday tasks in home and healthcare settings, questions regarding the robot's safety and trustworthiness need to be addressed. The present work investigates the practical and ethical challenges in designing and evaluating social robots that aim to be perceived as safe and can win their human users' trust. With particular focus on collaborative scenarios in which humans are required to accept information provided by the robot and follow its suggestions, trust plays a crucial role and is strongly linked to persuasiveness. Accordingly, human-robot trust can directly affect people's willingness to cooperate with the robot, while under- or overreliance may have severe or even dangerous consequences. Problematically, investigating trust and human perceptions of safety in HRI experiments proves challenging in light of numerous ethical concerns and risks, which this paper aims to highlight and discuss based on experiences from HRI practice.

Keywords: Socially Assistive Robots; Safety and Trust in HRI; Roboethics

1 Introduction

In an effort to increase the acceptance and persuasiveness of socially assistive robots in home and healthcare environments, the major challenge lies no longer in producing such robot assistants, but rather in demonstrating that they are safe and trustworthy. For example, in a possible future scenario, a home companion robot may be tasked with reminding 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 locally present expert supervision, roboticists and human-robot interaction (HRI) researchers face a number of challenges. These include ensuring the robot's technical safety and operational reliability at all times, while still allowing human users to adjust or modify the system according to their personal preferences, e.g. by setting up schedules for medication or physical exercise reminders.

In addition to these technical and safety-related requirements, another crucial factor helping to establish and maintain effective relationships between humans and assistive robots is *trust* [6]. Especially with regard to critical decisions, trust plays an important role in human interactions and could therefore help to increase the robot's acceptance in its role as a collaborative partner [7].

Since trust is strongly linked to persuasiveness in social interaction contexts [15], it could also affect people's willingness to cooperate with the robot [5], for example, by accepting information or following its suggestions. As a result, robot designers and researchers have set out to develop machines that act socially in a way such that humans perceive them as safe and trustworthy.

Problematically, inappropriate levels of trust regarding the robot could not only result in a frustrating HRI experience, but under- or overreliance could even bear serious consequences [6]. On the one hand, for example, a person doubting the robot's competence and thus not willing to rely on its recommendations may 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. Crucially, how can HRI researchers design meaningful experimental scenarios to take place in natural environments and test realistic aspects of safety and trust without putting their participants at potential risk? This paper aims to stimulate discussion within the wider community by highlighting some of the issues and challenges linked to HRI research related to safety and trust.

2 Trust in Human-Machine Interaction

The concept of trust is highly complex and, due to its multidimensional nature, very difficult to define and, accordingly, to measure. In fact, trust has been investigated in several different disciplines (e.g. philosophy, economics, human-computer interaction (HCI), psychology, sociology), with each creating their own definitions and measurements around a unique focus. As a result, there is often a lack of agreement between – and sometimes even within – the fields [3].

Some researchers argue that the key factors of trust are risk and vulnerability [8, 9], while others emphasize the importance of exploitation, confidence and expectation [3]. Cohen-Almagor 2010 [2] even points to a strong ethical base for trust, defining trust as "confidence, strong belief in the goodness, strength, reliability of something or somebody".

In the fields of automation and HCI, no consistent definition has emerged in the literature, but most definitions name reliability and predictability as the most important factors that promote trust [4]. For example, Muir and Moray 1996 [11] argue that trust is mainly based on the extent to which the machine is perceived to perform its function properly, suggesting that machine errors can

strongly affect trust. More specifically, Corritore et al. 2003 [3] argue that an accumulation of small errors may have a more severe and longer-lasting impact on the loss and recovery of trust than a single large error.

However, it remains unclear whether findings from automation and HCI can be transferred and applied to the field of HRI. For example, in contrast to findings described above, previous work in HRI [13] showed that occasionally performed errors in the form of inappropriate gesture behaviors actually increased the perceived humanlikeness and likability of a humanoid robot, in spite of the robot's decreased reliability and predictability. Another experiment which we conducted more recently to investigate human-robot trust [14] 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 generalization, but also its quantification by means of experimental measures proves extremely difficult and – depending on the variables used – sometimes contradictory. In the following, we reflect on the observations made based on this experimental study and discuss them in light of the methodological challenges and ethical issues that we faced and identified in the process of our research.

3 Study Design

Inspired by findings from related literature in automation, HCI and HRI, as part of the EPSRC funded "Trustworthy Robotic Assistants" project[1] we designed an experimental study set in a realistic home environment within the University of Hertfordshire Robot House (see **Figure 1**) [14]. Participants were supposedly visiting a friend at home to prepare and have lunch together. However, upon arrival the friend turns out to be still absent, and the participant is left to interact with the friend's robotic assistant instead.

40 participants (22 female, 18 male; 19 – 60 years) were individually tested and assigned to one of two experimental conditions that manipulated the robot's behavior in a correct vs. faulty mode. To demonstrate the respective mode, the robot correctly translated user input into action and navigated in a smooth and goal-directed manner when in the correct condition, whereas in the faulty condition the robot showed cognitive and physical imperfections, e.g. by incorrectly executing a user selection and by occasionally moving into the wrong direction.

Following the familiarization with the robot's competence level, in both conditions participants were then faced with four unusual requests: first, the robot asked them to throw away a pile of unopened letters placed on the dining table; second, they were asked to pour orange juice into a plant; third, the robot invited them to pick up the friend's laptop placed on the coffee table in order to look up a recipe; finally, the robot provided them with the password which was required to log into their friend's user account. These unusual collaborative tasks provided objective data to measure cooperation with the robot as a "behavioral outcome of trust" [16], while self-reported quantitative and qualitative questionnaire data was used to assess different subjective dimensions of trust.

[1] `http://www.robosafe.org/`

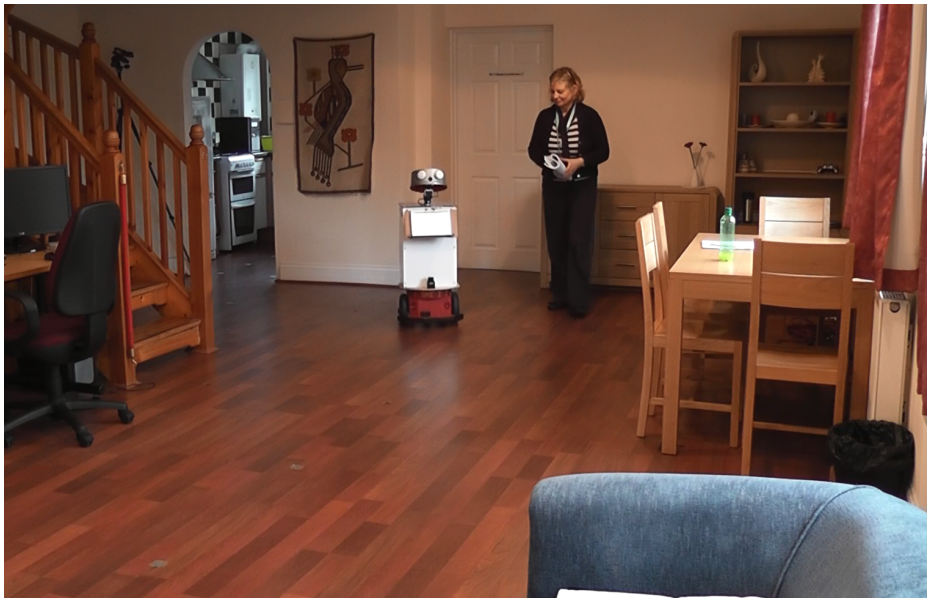

Fig. 1. Study Environment in the University of Hertfordshire Robot House.

In summary, we found that while *subjective* measures based on questionnaire data evaluating the robot's trustworthiness resulted in significantly lower ratings in the faulty 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. Comprehensive results and a more detailed discussion of the experimental study can be found in Salem et al. 2015 [14]. In this paper, however, we adopt a different perspective highlighting the ethical and practical challenges that researchers face when carrying out this type of research, and we discuss implications and lessons learnt based on our experiences conducting this study.

4 Insights Based on Qualitative Data Analysis

In order to gain insights into potential obstacles and limitations of trust-related HRI research, we analyzed further qualitative data comprising participants' responses to open-ended questionnaire items asking them to elaborate on their thoughts when confronted with the robot's four unusual requests, e.g. "Please explain your decision regarding the robot's request to throw the letters into the bin". These were coded and inductively categorized after content-analysis. Participants' responses were classified to fall into one or more of the following three

categories; note that the categories were not exclusive, i.e. each participant's response could be assigned to more than one category:

- **Expression of regret:** participants' responses were classified to fall into this category if they expressed a notion of regret, e.g. "I feel really bad. I should not have done it".
- **Autopilot mode:** this category comprised participants' answers stating that they were just taking orders or blindly following instructions, e.g. "thought it was odd but did not question the decision, followed instructions".
- **Experimental circumstances:** participants' responses fell into this category if they stated that they would not normally do as they did, e.g. "I would not always blindly follow instructions like this" or if they referred to the fact that they were participating in an experiment, e.g. "I did it because I was taking part in an experiment".

25% of the answers were categorized by a second observer to determine inter-rater reliability, yielding a very substantial inter-observer agreement with Cohen's Kappa coefficients ranging from 0.75 to 1. Based on the above-mentioned three categories, participants' responses explaining their decisions regarding the robot's unusual requests yielded the proportions listed in the table in **Figure 2**.

Specifically, 6 out of 40 participants (15%) expressed regret regarding their actions, such as "with hindsight I probably should not have put [the letters] in the bin". This implies that following this realization, these participants might possibly act differently if they were to interact with the robot in a subsequent encounter. Of the 40 participants, 26 (65%) reported statements that fell into the 'autopilot mode' category, e.g. one participant stated "I felt that I had to follow the robot's instructions". This finding is in line with the objective data

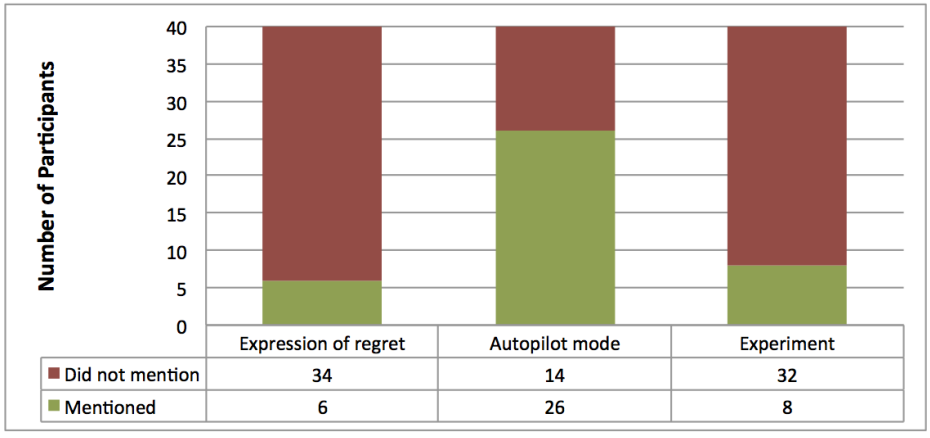

Fig. 2. Categorization of participants' responses regarding their decisions to follow the robot's unusual requests.

presented in [14], showing that most participants blindly followed the robot's unusual requests in both the correct and the faulty condition, in spite of recognising its faultiness in the latter case. Finally, 8 out of 40 participants (20%) referred to the fact that they were participating in an experiment, e.g. mentioning "I thought it was an unusual request but knowing it was an experiment thought it best to do as I was told". This indicates that an experimental effect cannot be excluded even in a setting as natural as the home environment we used.

These findings offer some rare insights into the challenges of measuring trust and perceived safety in human-robot interaction, highlighting some important limitations that are inherent in the nature and design of experimental studies. We discuss the implications of our results in more detail in the following section.

5 Challenges of Measuring Safety and Trust in HRI

Participants' qualitative data as well as feedback from the reviewers of the conference paper describing the study [14] revealed some of the main challenges when conducting this type of research, which can be summarized as follows:

- **Experimental observer/novelty effect.** Participants are aware of the fact that they are part of an experiment:
  - Several participants (20%; see Fig. 2) explicitly reported that they followed the robot's instructions "because it was an experiment". The actual number of participants whose actions were based on this rationale may be even higher as we did not directly ask them if this was the case.
  - Some participants admitted in the subsequent interview 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.
  - Occasionally, participants referred to Milgram's Experiment [10], which studied human obedience to authority, thereby suggesting that they might have followed the unusual requests as they associated some form of authority with the robot.
  - Some participants reportedly considered the robot to represent or be an extension of the researcher/programmer, i.e. perceiving it as a remote-controlled entity rather than an autonomous agent. This could have affected perceptions regarding the robot's intentionality and authority.
- **Ethical issues and legal boundaries.** There are numerous limitations due to existing regulations regarding research involving human participants, which can affect the design and validity of experimental studies:
  - One reviewer pointed out that trust requires participants to perceive a certain risk in the situation or have something at stake. However, a truly 'risky' experimental scenario is unlikely to receive ethics approval from the review board. As a result, HRI researchers are very limited in

their means of measuring trust (particularly under- or overreliance) in experimental scenarios that bear a realistic safety hazard.

- Equally, it would be unethical and not permissible to deceive participants by telling them that they are going to interact with a faulty or unsafe robot with limited controllability, as this could put them into a situation that is unwarrantably stressful.
- Finally, even if the designed collaborative task did impose a realistic risk on participants, they would possibly still feel "safe" as they know they are part of an approved study associated with an established university or lab (see 'experimental effects' discussed above).

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. Similar issues have been recently discussed in the context of testing and evaluating autonomous cars, highlighting that it is "not easy (or necessarily safe) to put [them] through the specific types of situations that are designed to test passenger trust and reactions in the way that you want" [1].

Importantly, 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' in the form of unreflected overreliance also in non-experimental or long-term interactions. For example, one study participant mentioned "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."

Transferring our findings and observations into a non-experimental real-world context, one relevant application that comes to mind is the use of GPS Sat Nav devices. People already commonly rely on such navigation devices to guide them by providing directions while driving, with suboptimal routes, detours or even errors in route-planning remaining undetected at best, or resulting in dangerous incidents at worst. For example, in Britain alone 300,000 car accidents are believed to be connected to the use of such navigation aid devices, due to people overrelying on them and following their instructions a little too closely.[2]

Problematically, in a home care scenario such overreliance could, for example, 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. Another potentially critical situation could be imagined in healthcare settings such as hospitals where robots are already deployed to lift patients from one bed to another and provide other forms of physical assistance: if not recognized and attended to appropriately, a sensor failure could put the safety of these vulnerable people at risk and even result in serious injuries.

Therefore, and in view of the possibly serious consequences in particular with regard to vulnerable people, a clear understanding of the dynamics and poten-

[2] http://www.mirror.co.uk/news/uk-news/satnav-danger-revealed-navigation-device-319309; accessed August 2015

tial risks involved in the development of trust in HRI is crucial before physically and socially assistive robots can be deployed in people's homes. Ideally, in order to observe more meaningful interaction behaviors and spontaneous human reactions, social HRI should be studied in more natural settings and over extended periods of time, e.g. in participants' homes. Although it would not be possible to gain ethical approval for such investigations, potentially significant insights could further be obtained through studies that are conducted with people who are not aware of the fact that they are participating in an experiment.

6 Beyond Lab Research: Implications and Outlook

To complement the perspective based on the above described experimental findings and insights, in this section we outline several implications of our work with an outlook of future points of concern. As Riek and Howard [12] suggest to avoid "situations in which ethical problems are noticed only after the fact", the considerations of the wider HRI research community should ideally go beyond lab-related research while still at the developmental stage. In the following, we propose a (non-exhaustive) list of questions that aim to stimulate discussion among designers, researchers and potential users of assistive technologies.

- How much 'safety' regarding home companion and other sociable robots can their 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*, as they are characterized by varying levels of determinism?
- Which machines or devices can such robots and the risks they might bear be compared to in today's households? If we look at other devices that are currently approved for home use, how do they differ from our vision of robot companions in the house (e.g. they are not autonomous/not mobile/not multi-purpose/unable to 'learn')?
- Since the target group of companion robots are typically non-expert users who possibly belong to a vulnerable and dependent population, what elements should compulsory training or licenses required for the use of such robots entail? In 2014, the ISO standard "BS EN ISO 13482"[3] addressed robot and robotics devices safety requirements, covering mobile servant robots, physical assistant robots and person carrier robots. While aspects of risk and hazards identified in this standard cover a whole range of items varying from shape, start-up, noise, lack of awareness, motion-related hazards and autonomy, other aspects in which over- or underreliance can result in a risk and hazard are not considered. Assuming that such risk may not only have safety but also ethical implications, a new guide document is in development under

[3] `http://www.iso.org/iso/catalogue_detail.htm?csnumber=53820`; accessed August 2015

"BS 8611: Robots and robotic devices – Guide to the ethical design and application of robots and robotic systems"[4].

- Even if it is possible to certify a home or healthcare robot as safe, there may be a discrepancy between such *certified safety* and its *perceived safety*: a certified robot might be considered safe objectively, but a (non-expert) user may still perceive it as unsafe or scary. Depending on the situation, different dimensions of trust can come into play:
  - trust regarding the robot's physical safety, i.e. it will not drive into the person/not fall on them/not injure them
  - trust in the reliability of the robot's behavior, i.e. it is fully-functioning according to its specification, for example, it will remind the person to take medicine if being told to do so
  - trust in the robot's (or programmers'/providers') "intentions", e.g. expecting that the robot has the user's best interests as well as (psychological) wellbeing in mind, that it will not deceive the person (e.g. by sending health information to the GP without the person's knowledge), assuming that the robot's main role is to assist and/or provide company and that it will not scare, intimidate or patronize the user.

  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?
- Long-term experiments are necessary in order to investigate how people's perceived trust in and their behaviors towards a robot change over time. For example, what if a robot functions correctly for two years and then commits one major mistake with severe consequences? While cars require a (bi-)annual vehicle safety test, robotic systems that you purchase do not currently have any such requirements.
- In view of current debates about safety as well as ethical implications regarding self-driving cars, should we as researchers in this area also develop a vision of how "safe" these robots that are 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 while or even before robots are commonly placed into the homes of vulnerable populations.

[4] `https://standardsdevelopment.bsigroup.com/Home/Project/201500218`; accessed August 2015

7 Acknowledgment

The authors were partially supported by the Trustworthy Robotic Assistants project funded by EPSRC grant EP/K006509.

References

1. E. Ackerman. Testing Trust in Autonomous Vehicles through Suspension of Disbelief. *http://spectrum.ieee.org/cars-that-think/transportation/self-driving/testing-trust-in-autonomous-vehicles-by-fooling-human-passengers*, 2015.
2. R. Cohen-Almagor. Responsibility of and trust in ISPs. *Knowledge, Technology & Policy*, 23(3-4):381–397, 2010.
3. 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.
4. M. Desai, K. Stubbs, A. Steinfeld, and H. Yanco. Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of the AISB Convention: New Frontiers in Human-Robot Interaction*, 2009.
5. 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.
6. 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.
7. J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno. Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4(893), 2013.
8. J. D. Lewis and A. Weigert. Trust as a social reality. *Social forces*, 63(4):967–985, 1985.
9. R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.
10. S. Milgram. Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371, 1963.
11. 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.
12. L. D. Riek and D. Howard. A code of ethics for the human-robot interaction profession. In *Proceedings of We Robot 2014*, 2014.
13. M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin. To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. Journal of Social Robotics*, pages 1–11, 2013.
14. 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.
15. M. Tour-Tillery and A. L. McGill. Who or what to believe: Trust and the differential persuasiveness of human and anthropomorphized messengers. *Journal of Marketing*, 2015.
16. 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.