ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

DEMONSTRATOR PROJECT

Final report

Assuring safety and social credibility

JULY 2019

Project report: Assuring Safety and Social Credibility

Dr Catherine Menon (c.menon@herts.ac.uk)

Dr Patrick Holthaus (p.holthaus@herts.ac.uk)

Professor Farshid Amirabdollahian (f.amirabdollahian2@herts.ac.uk)

This work was supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Registry Foundation and the University of York

1. Introduction

Assistive robots offer significant benefits to an increasingly elderly population, both in terms of their social impact and their functionality (Broekens, 2009), (Amirabdollahian, 2013). Assistive robots support independent living by aiding humans to conduct basic activities, such as preparing food and bathing. Similarly, these robots may support the psychological health of elderly or isolated individuals via socially-important behaviours, providing companionship and encouraging these individuals to engage and interact.

In order to be effective as assistive devices, these robots must perform functions important to safety, such as alerting a user if an appliance has been left on, alerting a user if medication has not been taken, or encouraging a user to perform necessary rehabilitative or medical actions. Simultaneously, if these robots are to be effective assistive devices they must also demonstrate behaviour which is sufficiently socially acceptable for the end-user to fully engage with them.

As a foundation for this project, we have postulated that the following question represents a critical barrier (C-BAR):

Social acceptability: where user acceptance of an RAS depends on effective performance of social functions, how can potentially conflicting social and safety requirements be balanced and how can we assure that the RAS is both safe and acceptable to end users?

1.1 Project work and results

In this six-month project we have identified and characterised a link between social credibility and effective performance of safety-related behaviours. We have demonstrated the presence of this link in an experimental domestic environment setting, showing that an assistive robot which does not perform adequate social behaviours is also less effective at performing safety-related behaviours in the home. We have identified the types of social behaviour which have an increased effect on safety performance, and considered situations where the social and safety requirements might conflict. Although further work must be done to generalise this to other situations and autonomous systems, the experimental results validate the C-BAR identified above, and indicate some methods of overcoming it. These are discussed further in Section 4.

2 Introductory work and hypothesis

Our introductory work for this project (Menon, 2019) was presented in March 2019 at the 9th International Conference on Performance, Safety and Robustness in Complex Systems and Applications, and was awarded a Best Paper prize. In this paper we performed a literature survey which identified that the social effects of assistive robots are not typically factored into hazard analysis, and equally, that there is often very little consideration of the ways in which the social performance of an assistive robot are affected by safety features (e.g., automatic stops, avoidance of physical contact). This paper served as a foundation for discussion on how to bring these concerns together within a single domain, and more widely, how to assure the safety of an autonomous system (from any domain, including automotive and health) which must also perform another social function.

In this paper we examined how both the safety-critical and socially important behaviours of an assistive robot rely on the user's engagement with the robot. We considered the Care-O-Bot (Kittmann, 2015), which is an up-to-date example of a mobile assistant robot with the capacity for social interaction. The Care-O-Bot is typically expected to perform a range of functions including:

- Alerting a user if an electrical appliance is malfunctioning
- Reminding a user to take their medication
- Reminding a user if an appliance has been left on

In addition to these care-giving behaviours, the Care-O-Bot is expected to encourage the user to engage and interact by offering entertainment and companionship.

Some of these functions have the potential to impact safety. The robot presents both physical hazards (e.g., its weight can contribute to crush injuries) as well as functional hazards. For example, the robot may fail to perform a safety-critical function such as reminding the user to take medication or may perform this function incorrectly (e.g., reminding the user too frequently). A further concern with the Care-O-Bot is the ability for end-users to define their own desired robot behaviours. This is important for user engagement – for example, a user may wish the robot to greet them as they enter the house – but a potential concern for safety. Specifically, there is the potential for an inexperienced end-user to define behaviours which impact safety, or which put the robot in a position which can violate assumptions about the constraints it will obey. Equally, a user may define a behaviour which causes the robot to remain in another room, compromising its availability to perform those safety-critical functions which rely on direct observation of the user.

The Care-O-Bot, like all assistive robots, is designed with reablement as a priority (Amirabdollahian, 2013). Reablement is defined as the drive to ``Support people to do rather than doing to / for people'' and is an important characteristic for service and assistive robots. Designing with reablement in mind means that the assistive robot is not intended to carry out the tasks itself (e.g., administering medicine to a user), but is instead intended to encourage the user to complete the task themselves. A direct result of this design principle is that the assistive robot will alert the human user to a potential hazard (e.g. an appliance that has been left on) but the user must take action themselves to complete the mitigation of any safety risk. That is, the human user must switch off the oven, repair the appliance, take the medication etc. as necessary, rather than relying on the robot to do this. In this way, the mitigation of safety risks is split between the assistive robot and the human user, and hence the effectiveness of safety performance is directly related to the extent which a user is willing to engage with the robot.

In this paper we introduced the concept of social credibility, this being a measure of how well an assistive robot obeys the social norms relevant to its environment. These social norms will be specific to the environment, and for a domestic assistive robot may include:

- Frequency and urgency of any interruptions
- Nature and intensity of interaction, engagement and interruptions
- Responsiveness of the robot to verbal and non-verbal feedback
- Appropriate physical movement and distance maintained from end-user

Social credibility is an evolving measure, and dependent on the actions of the robot. Much as a system which does nothing is “perfectly safe", a robot which is turned off and hence never takes an action will not lose nor gain social credibility. Social credibility may be temporarily lost by an inappropriate action, and gained back by subsequent actions. A loss of social credibility (from any cause) can lead to an end-user disengaging with the robot, choosing either to ignore its prompts or to switch it off. Studies have shown that users are more willing to switch off robots if they consider them to be solely robotic devices, instead of intelligent, social beings (Bartneck, 2009). This is exacerbated when the mode of engagement with this robot becomes arduous. In (Wall, 2013), drivers concluded that they would prefer to be able to turn off a speed warning system that was judged irritating, even where they agreed that use of the technology would be helpful.

Because assistive robots, by design, rely on the end user to complete mitigation of an identified safety risk, user disengagement compromises the ability of these robots to perform their safety-critical functions. For example, a robot reminding the user that the oven has been left on has no effect unless the user engages with the robot, and moves to switch the oven off. This is particularly true where the robot alert contradicts some existing mental model that the user has of the environment (“I didn’t leave the oven on”). In the aviation domain – where autonomous cockpit systems are not considered to be social entities – pilots have been observed to attempt to debug the automation when its actions

deviate from those they expected (Sarter, 1997). This is also a risk for assistive robots when the end user considers them to be monitoring devices only, instead of social, intelligent beings.

In this paper we suggest a number of potential methods to address loss of safety-critical functionality resulting from lowered social credibility. Each of these methods trades a slight decrease in the robot's overall capability in return for maintaining an adequate level of social credibility. Since social credibility is a requirement for effective safety critical performance, this corresponds to decreasing the robot's capabilities in order to gain confidence that safety-critical engagements will be performed effectively when needed. We therefore suggest that when the social credibility drops below a threshold value (termed the disengagement threshold), the robot alters the nature of its alerts and reminders to stop social credibility loss. For example, the robot may identify those alerts which are not safety-critical and choose to

- Avoid performing the alert entirely
- Delay the alert or perform it less frequently
- Slow its physical movements when coming to interrupt a user
- Decrease the volume of any audible alerts

This proposal allows a robot to omit routine behaviours (such as interrupting the user with the offer of food or drink) in order to retain sufficient social credibility to ensure that any safety-critical behaviour (such as notification the oven is on) will be engaged with by the user.

As a complicating factor, we also note that safety-critical performance is not the only consideration for assistive robots, and that such systems must also perform their social functionality adequately. There is the potential for prioritisation of functionality relating to safety (e.g., requiring the robot to follow the user through the house in case of a fall) to result in the neglect of other socially important behaviours such as greeting, user engagement and user interaction. In other words, a robot performing only safety-related behaviours may not be free to perform other roles which are critical to its reablement functionality.

3. Experimental work and results

The second major aim of this project was to perform an experiment that would validate the hypothesised link between social credibility and safety. We conducted a preliminary study with 30 participants that investigates their responses when notified of different environmental hazards by either a socially credible robot, or a robot that explicitly violates social norms. The study was carried out in the Robot House, a four-bedroom home used by the University of Hertfordshire for human-robot experiments. It is equipped with standard furniture and appliances, as well as smart home sensors and actuators. The experiment was approved by the University of Hertfordshire's Health, Science, Engineering and Technology Ethics Committee under protocol number COM/SF/UH/03714.

Participants were instructed to sit at a table in the Robot House, and complete as many cognitive tasks (i.e., Sudoku puzzles) as possible in the allotted time. This condition allowed us to simulate a valid, cognitively-engaging task which a user might be involved with when interrupted by a robot. Participants were informed that the robot may interrupt them at times during this task, and that it was their choice whether or not to perform an action in response to this interruption.

During the experiment, all participants were interrupted four times as follows:

- The robot informed them the oven in the kitchen was left on
- The robot informed them the power sockets in the kitchen were on
- The robot informed them some of the power sockets in the kitchen were still on
- The robot informed them a Pepper Robot in a different room was overheating while charging

Of the 30 participants, 15 (chosen randomly) worked with a robot which violated social norms (VN), and 15 with a robot which complied with social norms (AN). This condition was determined by the following robot behaviours:

- Distance during initial robot greeting (appropriate vs too far)
- Passing distance when moving or prior to interruption (appropriate vs too close)
- Position during interruption (frontal vs from behind)
- Head position during interruption (facing the participant vs facing the floor)
- Verbal utterances (abrupt vs polite)

No other condition was varied other than these, and the robot moved about the house in between interruptions to simulate a working domestic environment. As part of the experimental set up, we ensured that the oven in the kitchen was demonstrably left on. The power switches were also demonstrably on, and were turned on repeatedly via the house's smart sensors during the experiment. The Pepper Robot was visibly charging in a different room, but its display screen was deliberately set up to make it difficult for a participant to verify whether it was overheating or not.

3.1 Measurements

During the experiment we observed the participants via camera feed and smart sensors, and made objective measurements of:

- Physical response to interruptions (e.g. standing up)
- Movement made in response to interruptions (e.g. going into the kitchen)
- Extent of action taken to eliminate the hazard (e.g. switching off one or multiple power sockets)
- Time taken to perform an action that eliminates the hazard

3.2 Questionnaires

Following the experiment, participants were asked to complete questionnaires to ascertain their impression of robot's social behaviours (Carpinella, 2017), (Bartneck, 2009). These are established, validated questionnaires which are used extensively in HRI studies to assess the social, trust and emotional responses that users have to robots. In addition, we also asked each participant to:

- Rate their assessment of the severity of each hazard (oven, plugs and Pepper Robot)
- Rate their willingness and thoroughness to react to robot warnings
- Explain why they decided to act or not act in response to each interruption

3.3 Results

As this study was a preliminary study, and consisted of only 30 participants, no statistical significance between conditions was expected for the evaluation of the questionnaires. However, we were able to identify a number of trends from the collected data. We have summarised the trends here, and the quantitative results – including graphs and statistical analysis – are available in (Menon, 2019b).

3.3.1 Questionnaire results

The questionnaires about social perception and emotional response to the robot (Carpinella, 2017), (Bartneck, 2009) demonstrated clearly that the participants in the AN condition-set considered the robot much more socially credible than participants in the VN condition-set. The AN condition-set scored the robot higher on "positive" attributes such as sociability, responsiveness, competency, intelligence and consciousness while simultaneously scoring it lower on "negative" attributes such as aggression, strangeness and awkwardness.

The questionnaires about perception of hazards demonstrate that users considered the oven to be the most safety-critical hazard, followed by the overheating Pepper Robot, then the kitchen power

sockets. Over all these hazards, participants from the AN condition-set consistently rated these as more dangerous to safety than participants from the VN condition-set.

3.3.2 Measurable results

The AN condition-set participants were overall more likely to respond to the robot's interruptions than the VN condition-set participants for the oven hazard (79% vs 50%), the second power socket warning (71% vs 31%) and the Pepper Robot warning (79% vs 56%).

In addition, when AN participants responded to the robot, they were much more likely to take actions which corresponded to mitigating the hazard (e.g. turning the oven off). By contrast, when VN participants responded to the robot, their actions in many cases were observational only: 15% – 20% of VN participants chose simply to examine the environment without taking further action to mitigate the hazard. When VN participants did mitigate the hazard, they took longer on average to do this than the AN participants, and visually assessed the environment more thoroughly before doing so.

One of the most notable results was in the second power socket warning, where the biggest difference between AN and VN participant results was observed. This interruption elicited the lowest response rate (31%) from VN participants, while the response rate from AN participants remained high at 71%.

4. Further discussion

Although the small sample size means that statistical significance would not be expected, the identified trends provide some indication of how safety assurance might be affected by an autonomous system's social behaviours in this domain.

The most notable impact is on the user's willingness to accept the robot's assessment of hazards and the extent to which the user considers it necessary to "cross-check" these against their own experience. AN participants were more likely to respond at all to the robot's interruption, more likely to take an action that would mitigate the hazard, and more likely to perform this action quickly. By contrast, VN participants were less likely to respond at all, and more likely to spend time visually assessing the environment before taking any action. When asked about this in the open-ended questionnaires, VN users indicated that they were checking to see if the robot was "right" about the existence of the hazard.

This behaviour of disbelief was seen most clearly with the second power socket warning. This warning took place after the users have already been warned about the power sockets once (and given the opportunity to switch them off). As part of the experiment we remotely switched the sockets back on, without telling the participants. Although participants were aware that this was within the technical capabilities of the house, this still corresponded to a situation where the robot was informing them of a hazard which they had already mitigated. This gave rise to the biggest difference in the study between AN and VN participants: only 31% of VN participants responded, whereas 71% of AN participants did. This demonstrates that AN participants are more likely to accept the robot's assessment of a situation, even where this directly contradicts their own experience.

This general effect can be taken to indicate that when it comes to assessment of safety-critical situations, users are more likely to believe a robot that they consider socially intelligent instead of one lacking social competency. This models the effect seen in emergency evacuations, where people are more likely to respond to warning messages delivered by a human instructions than to those delivered by an automated warning system (Mileti, 1990). Identifying that this effect also holds true for socially competent autonomous systems is a significant step in considering the effectiveness – and therefore the safety assurance – of such systems.

We also note that this effect is lasting, with the users' post hoc assessment of hazards also being dependent on the robot's social competence. The questionnaires on hazard severity identified that all

hazards (oven, power sockets and overheating robot) were consistently rated as more dangerous to safety by AN participants than VN participants. This gives rise to some interesting questions about the extent to which the perceived safety of other systems might be affected by previous interactions with a social robot. Such considerations are important for Systems of Systems (SoS) in which the assurance of one component of a system might be dependent on another, or where a single operator is responsible for interactions with multiple components.

We do also note a confounding factor, which is where social and safety behaviours are at odds with each other. Section 2 considers this from the point of view of interruptions (more interruptions to alert to hazards may increase safety, but at the risk of diminishing the robot's safety credibility) but it is also a potential factor in other interaction details. In particular, participants in the AN condition were interrupted by a robot that was facing them, whereas VN participants were interrupted by a robot behind them. This meant that AN participants had to move closer to the robot in order to take action (i.e. they had to walk past it instead of away from it). While this interaction mimics a human interruption – and is therefore more socially credible – it also slightly increases the chance of a human-robot collision. Design of social autonomous systems must therefore carefully consider which social requirements might also introduce an unacceptable level of safety risk, and whether this might be mitigated by the "protective" effect of enhanced social credibility.

5. Future work

This project has verified that our hypothesised link between social credibility and safety exists, albeit with a small sample size. We would look to build on this by widening the research to consider more general assistive robots, as well as other autonomous systems which may be expected to perform safety-critical behaviours alongside others (e.g. social behaviours, security behaviours). We anticipate that this work will take place in three phases as follows.

5.1 Phase 1: Behaviour identification

We propose to begin by identifying the most common or pervasive social and safety critical behaviours performed by assistive robots. This generalises from the constrained environment of our initial experiment to provide real-world results based on existing research (Syrdal, 2009), (Saunders, 2016) (International Standards Organisation, 2014), (Fong, 2003).

We will then determine which of these social behaviours are effective in building social credibility, and characterise the interactions between these behaviours and the safety-critical functionality of the robot. This will allow us to define a minimum necessary threshold for social credibility that also assures effective performance of safety-critical functionality. We will validate this by further experimentation, with particular focus on the areas in which social and safety behaviours conflict.

5.2 Phase 2: scheduling and prioritisation

We will follow this with a phase in which we identify scheduling and prioritisation mechanisms for an assistive robot. These mechanisms must take both safety requirements and social requirements into account in order to schedule behaviours and adjust the robot's functionality. These scheduling mechanisms must be able to dynamically change priorities based on the environment – e.g. the prevalence and severity of hazards – as well as the current social credibility which the robot has built up. We will build on existing work on safety in mixed-criticality systems (Iacovelli, 2018) to determine the constraints on these scheduling mechanisms which ensure that the residual risk of the robot is acceptable. The following (non-exhaustive) criteria are likely to affect the scheduling of robot behaviours and the nature of these behaviours:

- Estimated risk associated with not fulfilling the behaviour
- Estimated loss of social credibility associated with fulfilling the behaviour
- Current social credibility

- Functional importance of other behaviours

5.3 Phase 3: extension of results

In this phase we will seek to generalise our results from the domain of assistive robots to wider domains including health care and automotive. Autonomous systems in these domains may need to interact with the public, and thus perform a social (or at least, a societal) role. We will seek to define generalised assurance techniques for balancing safety with potentially conflicting social properties. We will extend the concept of scheduling behaviours to ensure social acceptability, and provide a worked case study the verify the applicability of our results in wider domains.

References

F. Amirabdollahian, R. op den Akke, S. Bedaf, et al., "Assistive technology design and development for acceptable robotics companions for ageing years.," Paladyn: Journal of Behavioral Robotics, vol. 4, no. 2, pp. 94–112, 2013.

C. Carpinella, A. Wyman, M. Perez, S. Stroessner. "The robotic social attributes scale (ROSAS): Development and validation", in Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 254-262, 2017.

C. Bartneck, T. Kanda, O. Mubin, et al., "Does the design of a robot influence its animacy and perceived intelligence?" International Journal of Social Robotics, vol. 1, no. 2, pp. 195–204, 2009.

J. Broekens, M. Heerink, H. Rosendal, et al., "Assistive social robots in elderly care: A review," Gerontechnology, vol. 8, no. 2, pp. 94–103, 2009.

T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," Robotics and autonomous systems, vol. 42, no. 3-4, pp. 143–166, 2003.

S. Iacovelli, R. Kirner, and C. Menon. "ATMP: An Adaptive Tolerance-based Mixed-criticality Protocol for Multi-core Systems". In Proceedings of the 13th International Symposium on Industrial Embedded Systems. 190 – 199, 2018.

International Standards Organization, "Robots and robotic devices – safety requirements for personal care robots," ISO 13482, 2014.

R. Kittmann, T. Fr¨ohlich, J. Sch¨afer, et al., "Let me introduce myself: I am care-o-bot 4, a gentleman robot," in Mensch und Computer 2015 – Proceedings, S. Diefenbach, N. Henze, and M. Pielot, Eds., Berlin: De Gruyter Oldenbourg, 2015, pp. 223–232.

C. Menon, P. Holthaus. "Does a loss of social credibility impact robot safety?", in Proceedings of the 9th International Conference on Performance, Safety and Robustness in Complex Systems, pp. 18 – 25, 2019.

P. Holthaus, C. Menon, F. Amirabdollahian. "How a Robot's Social Credibility Affects Safety Performance?", *in preparation.*

D. Mileti, J. Sorensen. "Communication of emergency public warnings: A social science perspective and state-of-the-art assessment", Department of Energy Technical Report, ORNL 6609 / DE 91 004981, 1990.

N. B. Sarter and D. D. Woods, "Team play with a powerful and independent agent: Operational experiences and automation surprises on the airbus a-20," Human Factors, vol. 39, no. 4, pp. 553–569, 1997.

J. Saunders, D. Syrdal, K. L. Koay, et al., "'Teach Me - Show Me' - End-user personalisation of a smart home and companion robot," IEEE Transactions on Human-Machine Systems, vol. 46, no. 1, pp. 27–40, 2016.

D. S. Syrdal, K. Dautenhahn, M. L.Walters, and K. L Koay. "Sharing Spaces with Robots in a Home Scenario – Anthropomorphic Attributions and their Effect on Proxemic Expectations and Evaluations in a Live HRI Trial." In AAAI Fall Symposium "AI in Eldercare: New Solutions to Old Problems". Washington, DC, USA, 116–123, 2009.

J. Wall, V. Cuenca, L. Creef, et al. "Attitudes and opinions towards intelligent speed adaptation," in Intelligent Vehicles Symposium Workshops (IV Workshops), 2013 IEEE, IEEE, 2013, pp. 37–42.

ASSURING
AUTONOMY
INTERNATIONAL PROGRAMME