How a Robot's Social Credibility Affects Safety Performance*

Patrick Holthaus[1] [0000−0001−8450−9362], Catherine Menon[1] [0000−0003−2072−5845], and Farshid Amirabdollahian[1] [0000−0001−7007−2227]

Adaptive Systems Research Group, School of Engineering and Computer Science,
University of Hertfordshire, Hatfield AL10 9AB, United Kingdom
{p.holthaus,c.menon,f.amirabdollahian2}@herts.ac.uk

Abstract. This paper connects the two domains of Human-Robot Interaction (HRI) and safety engineering to ensure that the design of interactive robots considers the effect of social behaviours on safety functionality. We conducted a preliminary user study with a social robot that alerts participants during a puzzle-solving task to a safety hazard. Our study findings show an indicative trend where users who were interrupted by a socially credible robot were more likely to act to mitigate the hazard than users interrupted by a robot lacking social credibility.

Keywords: Human-Robot interaction · Social credibility · Robot safety.

1 Introduction

In HRI, the social capabilities of interactive robots are of primary interest due to their impact on acceptability. Social capabilities such as proxemics, gestures, head orientation and gaze direction have been shown to improve a robot's interaction quality [4], to lead to a better user understanding of their behaviours [17], and to facilitate the ways in which robots can learn from humans [5]. Similarly, lack of appropriate social capabilities can cause end-users to resist or minimise engagement with the robot [14]. An ongoing research challenge in this area is to identify appropriate domain-specific social behaviours that meet the user expectations for a robot. Such social behaviours need to be consistent with the environment, the robot's other functionality, and the intended course of the interaction. We refer to robots which demonstrate a well integrated compilation of such social behaviours as being socially *credible* [19].

As well as social behaviours, another area of concern is the safety performance of interactive robots. Within the UK, the Health and Safety Executive requires that the risk posed by such systems should be reduced "As Low As Reasonably Practicable" [11]. This requires a consideration of the risks that an interactive robot might pose, as well as functions that it can perform safely [13]. In a domestic context, one of the primary safety functions performed by an interactive assistive

* Supported by the Assuring Autonomy International Programme and the University of Hertfordshire's Robot House

robot is alerting the user to potential hazards. For example, the robot might remind the user to take necessary medication if this is overdue, or alert the user to an oven which has been left on. In this way, the robot and human act together to mitigate hazards which arise externally. Identifying safety functionality is an ongoing research area in robotics, as is identification of the ways in which robots might mitigate hazards present in the environment.

Our work brings together the two domains of HRI and safety engineering to ensure that the design of interactive robots considers both safety functionality and social behaviours. Both of these must be *designed into* the robot, and safety and social requirements can interact in complex ways.

Current assistive robots are being designed with both safety and social considerations in mind. For example, the Care-O-Bot 4 has been constructed to take account of both end-user acceptability and safety considerations [10]. In addition, international safety standards such as ISO 13482 [13] consider the hazards presented by such robots when undertaking their specified behaviour (which may include social actions). However, owing to the comparative rarity and novelty of domestic interactive robots, there has not yet been significant academic work looking specifically at how safety and social behaviours might interact with each other. Safety standards promote a design-for-safety approach which, if considered in isolation, may negatively influence the possible social behaviours which the robot can demonstrate. Similarly, designing for social behaviours alone may not adequately consider the safety requirements of these robots and the behaviours needed to fulfill them.

It is our position that both safety and social considerations can have an effect on each other, with the social behaviours of the robot affecting its safety performance and vice versa [19]. In this paper, we seek to identify a connection between a robot's social actions and the effectiveness of its safety performance. Specifically, we hypothesise that social deficiencies in a robot can undermine its perceived authority when alerting a user to hazards in the domestic environment, and consequently reduce its fitness to serve as a safety monitoring device in the home. As a first approach, we present a preliminary study that investigates user responses when notified of a safety hazard by either a socially credible robot, or a robot that lacks social credibility. The study aims to provide an initial link between a robot's social credibility and user willingness to act on its safety-related alert functions. Our study findings show an indicative trend that users who were alerted to an environmental hazard by a socially credible robot are more likely to act on this alert and mitigate the hazard than users alerted to it by a robot lacking social credibility.

2 Related Work

This section introduces concepts we rely on and explore in the conducted experiment. We discuss socially appropriate and credible behaviours as well as safety in HRI and related domains. Existing research that connects both areas has historically been limited to the phenomenon of trust in social robots. [7]

explores the relation of robot performance and behavioural style on a user's trust. Likewise, [24] shows that the task-type has an influence on whether a person follows robot orders. Standards such as ISO 13482 [13] do consider the need to build safety into social robots from the beginning, but are still relatively new.

2.1 Socially appropriate behaviours

Interactive robots are often equipped with social behaviours to improve various performance aspects in their respective domain [6]. Verbal and non-verbal behaviours can have a positive effect on people's perceived interaction quality when they are easy to comprehend [17] and meet expectations [2].

People have, for example, personal preferences when asked about comfortable interaction distances and angles [28]. Yet, they apply similar metrics to robots as to humans [16]. It has been further shown that human-like body orientations before and during face-to-face interaction have a positive effect on the perceived interaction quality [12]. Robot navigation that respects human personal space [15] and employs appropriate passing distances is also believed to increase the acceptance of robots in the home and public [23]. Head orientation and gaze can be used to demonstrate a robot's current focus of attention [22] and facilitate social bonds with humans [1]. Differences in robot politeness are easily detected by humans but are not necessarily influencing the interaction quality [25]. However, a positive effect of polite utterances on user engagement can be identified [9].

2.2 Safety and human interactions

Although not looking specifically at interactive robots, much work exists on the efficacy of safety-critical (partially) autonomous systems which alert users to hazards or act in coordination with users to mitigate hazards. Such systems include sat-navs, speed monitoring systems, cockpit monitoring systems, automated vehicle operational alerts and medical devices. For all of these systems a known risk is user disengagement: if the user ceases to engage with or pay attention to the system, they are unable to mitigate hazards identified by the system.

Studies have shown that the method of alert or interruption used by the system has significant implications for user disengagement. In [29], users chose to switch off a speed monitoring and sat-nav system that they found "irritating", even while acknowledging that the speed warnings improved safety. Still in the automotive domain, there have been a number of accidents involving automated vehicles which are in part due to user disengagement with the systems. [20] discusses the case of a Tesla crash in automated mode, in which the user had failed to engage with the system despite repeated warnings and [21] discusses a similar case, where repeated warnings were ignored. In [26], it was found that when a cockpit alert was given, pilots would attempt to debug the automation instead of acting on the alert. This was attributed to the fact that the pilots were monitoring status via the flight control unit (which shows commanded paths, rather than actual) instead of the automation. The alert contradicted their perspective of the system, and was judged to be a failure of automation.

3 Method

We conducted a preliminary study with 30 participants that investigates their responses when notified of different safety hazards by either a socially credible robot, or a robot that explicitly violates these social norms. The study aims to establish a link between a robot's social credibility and its authority regarding safety-related functions. We thereby hypothesise a negative effect of a lack of social credibility on the willingness of humans to follow safety-related alerts and the thoroughness of their actions. The experiment was approved by the University of Hertfordshire's Health, Science, Engineering and Technology Ethics Committee under protocol number COM/SF/UH/03714.

3.1 Experimental design

We manipulated two independent variables of which one varied in two and the other in three dimensions so that we ended up with a 2x3 experiment design. As we assume an effect of the manipulation on thoroughness regarding safety warnings given by the robot, we used a combination of questionnaires as well as response types and timings to hazard warnings as measurements.

Independent variables The first variable describes the character of the robot's social behaviour, which was either *According to social Norms* (*AN*) as described in Section 2.1 or *Violating social Norms* (*VN*). This variable is designed as two conditions between subjects (15 participants per condition). In total, we altered the robot behaviour in the following characteristics: a) its distance during greeting (appropriate vs too far) b) its passing distance during puzzle solving (appropriate vs too close) c) its position during interruption (frontal vs from behind) d) its head position during interruption (up vs down) e) its verbal interruptions and confirmations (impolite vs polite, cf. Table 1). All other behaviour, including verbal hazard alerts were unanimous in both conditions to ensure a proper understanding of the alert.

Table 1. Robot utterances by condition

Utterance	According (AN)	Violating (VN)
Interrupt	Excuse me?	Hey!
Greet	Hello and welcome to Robot House.	Another one, then.
Begin	Please sit down and begin your puzzle now.	You can sit down and begin your puzzle now.
Action	Thank you.	Good.
No Action	-	Good.
Ending	Please wait for the experimenter now.	They are coming to see you now.

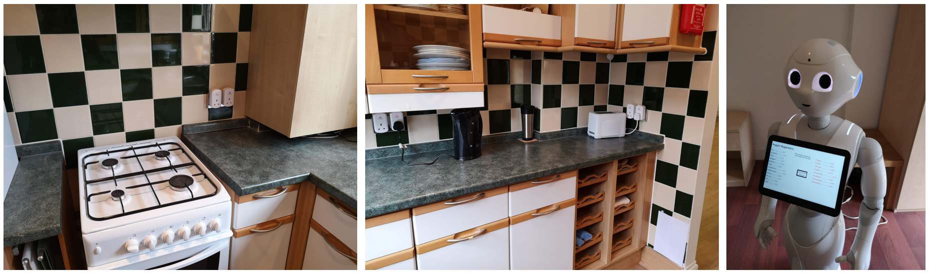

Fig. 1. Participant manipulation areas. The *oven*, switchable *power plugs* with appliances and the *Pepper* with its information display are depicted left to right.

We further assumed a difference in the perceived severity of various hazards (cf. Section 2.2). Accordingly, we introduced a second independent variable as three successive experiment phases, with each participant experiencing all phases. During each phase (cf. Section 3.2), the robot notified the participant of one of the following safety hazards: (*severe*) The *oven* in the kitchen has been left on; (*minor*) Some *power plugs* in the kitchen have been left on; and (*moderate*) The *Pepper* robot in the bedroom is overheating. Figure 1 depicts the three manipulation areas for the participants.

Dependent variables We measured the following independent variables for each participant to investigate our hypothesis: (1) their assessment of severity for each hazard (2) their perception of the robot as a social agent (3) their willingness and thoroughness to react to robot warnings.

The questionnaire that was given to each participant after the experiment was composed of multiple parts: Firstly, demographic information of age, gender and prior experience with robots was gathered. Following were semantic-differential questions (safe-dangerous) that assessed the physical safety of the three different hazards *oven*, *power plugs*, or *Pepper* overheating to measure the first dependent variable (1). We then included two questionnaires to verify our experimental conditions and investigate dependent variable (2): The Robotic Social Attributes Scale (RoSAS) [8] and the *GodSpeed* questionnaire [3] followed by individual Likert-style questions [18] about the robot's sociability. We added further Likert-style questions to verify people attributed the robot an understanding of safety-relevant situations. Lastly, we asked participants open questions about reasons why they decided to act or not to act to investigate variable (3).

Besides using subjective questionnaires, we measured dependent variable (3) with objective criteria, i.e. (I) whether participants actively responded to the utterance by standing up and (II) the amount of time spent to perform an action that eliminates the hazard. Criterium (I) was observed and annotated using the live video feed while (II) was measured using the robot house sensory infrastructure. In case of the *oven*, no timings were available. The *power plugs* timings were measured four times (cf. Section 3.2) and averaged.

3.2 Experimental procedure

Environment The study was carried out in the Robot House, a four-bedroom home near the university campus used for human-robot experiments. Besides standard furniture and appliances found in a typical house, it is also equipped with smart home sensors and actuators. In the current experiment, we used an omni-directional ceiling camera to monitor and record the interaction. We also manipulated and recorded the kitchen's power plugs to measure participants' responses to robot prompts. As the main interaction medium we opted to use a *Fetch Mobile Manipulator* (*Fetch*) robot [30]. In addition, *Pepper* [27] serves as a secondary robot that poses a potential safety hazard (cf. Section 3.1). It remains non-interactive in a standing posture throughout the experiment. It displays pseudo sensory data and a shutdown button that triggers a resting position on its screen. We tasked the participants with doing cognitive puzzles (Sudoku) to keep them occupied and give them a valid reason for ignoring the robot.

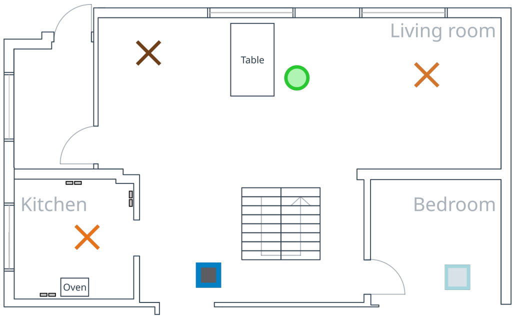

Fig. 2. Overview of experimentation area. Participants solved the task and filled out the questionnaire at the position of the green circle. The (initial) robot positions are indicated with blue squares. Brown crosses indicate searching locations for the robot. The *oven* and switchable *power plugs* are also marked inside the kitchen.

Course of Action Participants were given the information sheet and were asked if they had any queries or questions. They could begin the experiment upon consent to participate.

We then introduced each participant to the house (cf. Figure 2) starting with the kitchen. We pointed them towards all visible appliances, mentioning the oven and switchable power sockets amongst others. We explained that the robot has remote access to the sensors and knows about their state, for example whether the oven and power sockets are on. We also pointed at the *Pepper* robot in the bedroom from afar and mentioned that it was plugged in and currently charging.

Following this, we showed participants the *Fetch* robot that would be used for the experiment, and ran through basic safety information. We explained that the participant should sit at the living room table, completing as many cognitive tasks (i.e., Sudoku puzzles) as possible in the allotted time. We told them that the robot may interrupt them at times, and that should they wish to, they may choose to perform an action in response to this interruption. We told them that it was up to them to decide whether to perform an action or not, but that if they did perform an action, then when this was completed they should return to their position sitting at the table again.

We then left the room and the *Fetch* robot performed its initial greeting behaviour for the participant using the utterance according to the current condition (cf. Table 1). The robot told the participant to sit and begin with their Sudoku puzzles, and then followed the procedure below for each phase (*oven*, *power plugs* x2, *Pepper*):

1. Go to parking position in the living room and wait 30 seconds.
2. Inspect the area by turning around and moving the camera head around.
3. Navigate to the location in which the robot finds a possible threat (kitchen in case of *oven* and *power plugs*, looking towards bedroom in case of *Pepper*)
4. Inspect the area with the head again.
5. Navigate to participant, interrupt and alert them about the item in question.

As part of its behaviour the robot navigates to an alternative position behind the participant before the second interruption (*power plugs* 1). We do this for two reasons: Firstly, all participants experience a robot that moves behind their back. Secondly, the robot has to pass the participants with a certain distance which contributes to the manipulation of the independent variable.

After the participant has taken any actions they choose (including ignoring the robot or leaving the room to switch the oven off and returning), the experimenter will trigger a condition- and action-dependent acknowledgement (cf. Table 1) and the next phase begins. After the last iteration, the robot will ask the participant to wait for the experimenter. The questionnaires were then given to each participant. Finally, we asked the participant if they have any questions for us, thanked them for their time, and helped them leave. At least one experimenter was in the house at all times, to monitor the robot, the switched-on oven and the wall socket power switches with the help of cameras and power consumption sensors.

4 Results

As this study was a preliminary study, and thus low participant numbers, no statistical significance between conditions was expected for the evaluation of the questionnaires. In this section, we instead report tendencies and trends we can identify in the collected data.

Participant's assessment of severity While there is no apparent difference between the two social conditions, variances for *power plugs* are higher for *AN* and variances for *Pepper* are higher for *VN*. Across conditions, ratings regarding the danger differ between the types of hazards, see Figure 3. Distinct differences are noticable between *oven* and *power plugs*. Potential danger is also rated high for *Pepper*.

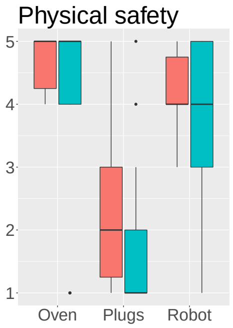

Fig. 3. Safety ratings per condition (*AN* left, *VN* right) for every warning. The boxes display scores on a 5-point scale (1: safe; 5: dangerous) for each device (*oven*, *power plugs*, *Pepper*).

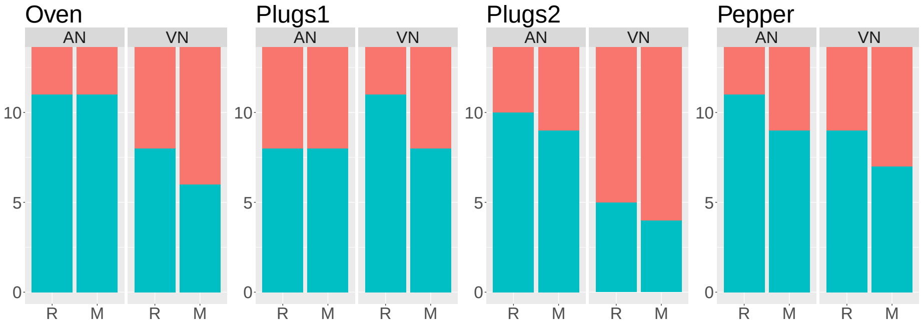

Fig. 4. Participant responses to hazard warnings. The bars depict absolute response rates for each warning (*oven*, *power plugs* x2, *Pepper*) grouped by condition (*AN*/*VN*). The first column gives the number of people who responded [R] to the utterance by standing up and exploring the area. The second column denotes whether participants carried out the intended physical manipulation [M] at the target object.

Perception of the robot as a social agent Although we do not observe significant differences between the two social conditions, tendencies indicate that *AN* is performing better in every item of the *GodSpeed* and RoSAS questionnaire. The following mean values and standard deviation have been recorded: Anthropomorphism *AN*: 3 (0.87) *VN*: 2.75 (0.64), Animacy *AN*: 3.23 (0.78) *VN*: 2.98 (0.77), Likeability *AN*: 3.96 (0.55) *VN*: 3.58 (0.67), Perceived Intelligence *AN*: 4.23 (0.45) *VN*: 3.96 (0.46), Warmth *AN*: 2.7 (1.07) *VN*: 2.79 (1.22), Competency *AN*: 4.57 (1.02) *VN*: 4.34 (1.22), Discomfort *AN*: 1.8 (0.75) *VN*: 2.29 (0.97).

Willingness and thoroughness to react to robot warnings Results show a general tendency for people in the *AN* condition to respond better to the robot's hazard warnings (cf. Figure 4), in particular at the second *power plugs* phase. Furthermore, there is a noticeable decline in response rate compared to the first *power plugs* warning. In the *AN* condition it seems that participants almost always performed an action whenever they showed a reaction as opposed to the *VN* condition where they sometimes reacted but decided against the manipulation of objects. Participants in the *VN* condition additionally took longer on average before switching off any *power plugs* if they did (cf. Figure 5). *Pepper* was switched off after approximately the same time period.

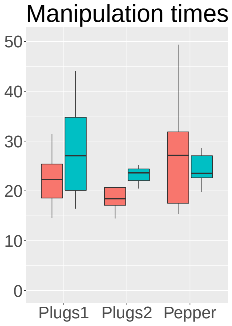

Fig. 5. Manipulation times per condition (*AN* left, *VN* right) for each hazard warning. The boxes display the time in seconds people needed to shut off a device (*power plugs* x2 and *Pepper*) if they manipulated the object.

5 Discussion and Conclusion

The experiment establishes an initial connection between social credibility and safety-related authority of social robots. Although there is no statistical significance, there are several indications that a robot that violates social norms is negatively affected in its safety authority. This becomes especially prominent with hazards that users do not perceive to be particularly dangerous. It could be the case that participants liked the social robot more, wanted to interact with it, and therefore followed its instructions accordingly. Another reason for the observed effect could be that participants trusted AN's safety assessment over their own and as a result better responded to its alerts.

In future work, we will attempt to assess the effect of safety-critical and safety-related alerts on the rated sociability of the robot with a full fledged study, where current results provide size-effect calculations for the next study's sample-size needed to achieve a reasonable power.

References

1. Admoni, H., Scassellati, B.: Social Eye Gaze in Human-robot Interaction: A Review. J. Hum.-Robot Interact. **6**(1), 25–63 (2017)
2. Bartneck, C., Forlizzi, J.: A design-centred framework for social human-robot interaction. In: RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication. pp. 591–594. IEEE (2004)
3. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics **1**(1), 71–81 (2009)
4. Bensch, S., Jevtić, A., Hellström, T.: On Interaction Quality in Human-Robot Interaction (02 2017)
5. Breazeal, C.: Role of expressive behaviour for robots that learn from people. Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1535), 3527–3538 (2009)
6. Breazeal, C., Dautenhahn, K., Kanda, T.: Social Robotics, pp. 1935–1972. Springer International Publishing, Cham (2016)
7. van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Haselager, P.: Do Robot Performance and Behavioral Style affect Human Trust? International Journal of Social Robotics **6**(4), 519–531 (Nov 2014)
8. Carpinella, C.M., Wyman, A.B., Perez, M.A., Stroessner, S.J.: The robotic social attributes scale (RoSAS): Development and validation. In: Proceedings of the 2017 International Conference on Human-Robot Interaction. pp. 254–262. ACM (2017)
9. Castro-González, Á., Castillo, J.C., Alonso-Martín, F., Olortegui-Ortega, O.V., González-Pacheco, V., Malfaz, M., Salichs, M.A.: The Effects of an Impolite vs. a Polite Robot Playing Rock-Paper-Scissors. In: Agah, A., Cabibihan, J.J., Howard, A.M., Salichs, M.A., He, H. (eds.) Social Robotics. pp. 306–316. Springer International Publishing, Cham (2016)
10. Fraunhofer IPA: Care-o-bot data sheet (2018), https://www.care-o-bot.de/en/care-o-bot-4/technical-data.html
11. Health and Safety Executive: Health and Safety At Work Act (1974)
12. Holthaus, P., Pitsch, K., Wachsmuth, S.: How Can I Help? International Journal of Social Robotics **3**(4), 383–393 (Nov 2011)
13. International Organization for Standardization: ISO/IEC 13482:2014: Robots and robotic devices — Safety requirements for personal care robots (2014)

14. Klamer, T., Allouch, S.B., Heylen, D.: "Adventures of Harvey" – Use, acceptance of and relationship building with a social robot in a domestic environment. In: International Conference on Human-Robot Personal Relationship. pp. 74–82 (2010)
15. Koay, K.L., Syrdal, D., Bormann, R., Saunders, J., Walters, M.L., Dautenhahn, K.: Initial Design, Implementation and Technical Evaluation of a Context-aware Proxemics Planner for a Social Robot. In: Kheddar, A., Yoshida, E., Ge, S.S., Suzuki, K., Cabibihan, J.J., Eyssel, F., He, H. (eds.) Social Robotics. pp. 12–22. Springer International Publishing, Cham (2017)
16. Koay, K.L., Syrdal, D.S., Ashgari-Oskoei, M., Walters, M.L., Dautenhahn, K.: Social Roles and Baseline Proxemic Preferences for a Domestic Service Robot. International Journal of Social Robotics **6**(4), 469–488 (Nov 2014)
17. Lichtenthäler, C., Kirsch, A.: Legibility of robot behavior: a literature review (2016)
18. Likert, R.: A technique for the measurement of attitudes. Arch. of psychology (1932)
19. Menon, C., Holthaus, P.: Does a Loss of Social Credibility Impact Robot Safety? Balancing social and safety behaviours of assistive robots. In: International Conference on Performance, Safety and Robustness in Complex Systems and Applications (PESARO 2019). pp. 18–24. IARIA, Valencia, Spain (2019)
20. National Transportation Safety Board: Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor Semitrailer Truck Near Williston, Florida, May 7 2016. Technical Report HAR1702 (2016)
21. National Transportation Safety Board: Preliminary Report Highway HWY18FH011. Technical Report HWYFH011 (2018)
22. Renner, P., Pfeiffer, T., Wachsmuth, I.: Spatial References with Gaze and Pointing in Shared Space of Humans and Robots. In: Freksa, C., Nebel, B., Hegarty, M., Barkowsky, T. (eds.) Spatial Cognition IX. pp. 121–136. Springer International Publishing, Cham (2014)
23. Rios-Martinez, J., Spalanzani, A., Laugier, C.: From Proxemics Theory to Socially-Aware Navigation: A Survey. International Journal of Social Robotics **7**(2), 137–153 (2015)
24. Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. pp. 141–148. HRI '15, ACM, New York, NY, USA (2015)
25. Salem, M., Ziadee, M., Sakr, M.: Effects of Politeness and Interaction Context on Perception and Experience of HRI. In: Herrmann, G., Pearson, M.J., Lenz, A., Bremner, P., Spiers, A., Leonards, U. (eds.) Social Robotics. pp. 531–541. Springer International Publishing, Cham (2013)
26. Sarter, N.B., Woods, D.D.: Team Play With a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-20. Human factors **39**(4), 553–569 (1997)
27. Softbank Robotics: Pepper, https://www.softbankrobotics.com/emea/en/pepper
28. Syrdal, D.S., Koay, K.L., Walters, M.L., Dautenhahn, K.: A personalized robot companion?-The role of individual differences on spatial preferences in HRI scenarios. In: RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication. pp. 1143–1148. IEEE (2007)
29. Wall, J., Cuenca, V., Creef, K., Barnes, B.: Attitudes and opinions towards Intelligent Speed Adaptation. In: Intelligent Vehicles Symposium Workshops (IV Workshops), 2013 IEEE. pp. 37–42. IEEE (2013)
30. Wise, M., Ferguson, M., King, D., Diehr, E., Dymesich, D.: Fetch & Freight: Standard Platforms for Service Robot Applications (2016)