---

**Article:**
Cameron, D. orcid.org/0000-0001-8923-5591, de Saille, S. orcid.org/0000-0002-8183-7771, Collins, E. et al. (5 more authors) (2021) The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. Computers in Human Behavior, 114. 106561. ISSN 0747-5632

https://doi.org/10.1016/j.chb.2020.106561

---

# The effect of social-cognitive recovery strategies on likability, capability and trust in social robots

David Cameron*[a], Stevienna de Saille[b], Emily Collins[c], Jonathan M. Aitken[d], Hugo Cheung[e], Adriel Chua[e], Ee Jing Loh[e], & James Law[f]

d.s.cameron@sheffield.ac.uk*, s.desaille@sheffield.ac.uk, e.c.collins@liverpool.ac.uk, jonathan.aitken@sheffield.ac.uk, mcheung3@sheffield.ac.uk, dxachua1@sheffield.ac.uk, ejloh2@sheffield.ac.uk, j.law@sheffield.ac.uk.

[a]Information School, The University of Sheffield, Sheffield, UK

[b]Institute for the Study of the Human, The University of Sheffield, Sheffield, UK

[c]Department of Computer Science, The University of Liverpool, Liverpool, UK

[d]Department for Automatic Control & Systems Engineering, The University of Sheffield, Sheffield, UK

[e]Department of Psychology, The University of Sheffield, Sheffield, UK

[f]Sheffield Robotics, The University of Sheffield, Sheffield, UK

Funding: This work is supported by the European Union Seventh Framework Programme (FP7-ICT-2013-10) under grant agreement no. 611971.

Declaration of Interest: None

*Corresponding author.

Abstract

As robots become more prevalent, particularly in complex public and domestic settings, they will be increasingly challenged by dynamic situations that could result in performance errors. Such errors can have a harmful impact on a user's trust and confidence in the technology, potentially reducing use and preventing full realisation of its benefits. A potential countermeasure, based on social psychological concepts of trust, is for robots to demonstrate self-awareness and ownership of their mistakes to mitigate the impact of errors and increase users' affinity towards the robot. We describe an experiment examining 326 people's perceptions of a mobile guide robot that employs synthetic social behaviours to elicit trust in its use after error. We find that a robot that identifies its mistake, and communicates its intention to rectify the situation, is considered by observers to be more capable than one that simply apologises for its mistake. However, the latter is considered more likeable and, uniquely, increases people's intention to use the robot. These outcomes highlight that the complex and multifaceted nature of trust in human-robot interaction may extend beyond established approaches considering robots' capability in performance and indicate that social cognitive models are valuable in developing trustworthy synthetic social agents.

*Keywords:* Human-Robot Interaction, Social Robotics, Trust, User Experience

# The effect of social-cognitive recovery strategies on likability, capability and trust in social robots

## Introduction

The emergence of social robotics has substantially expanded the potential roles that robots may have in society. Whilst existing autonomous robotic applications have kept robots largely segregated from humans (such as in manufacturing, warehouse logistics, and hazardous environments), there is a rapidly increasing market for robots in more public settings (e.g., entertainment, education, health and social care, domestic service, delivery and transportation, IFR Statistical Department, 2019). In these settings, robots will be expected to engage in reciprocal social interactions and promote cooperation from potential users (Lee, Kiesler, & Forlizzi, 2010). While robots may be able to perform useful actions in these flexible, social environments, whether people accept and intend to interact with them remains a key issue (e.g., Savela, Turja, & Oksanen, 2018). Given the widely reported evidence that people import social rules from human-human interactions to interactions with non-living agents (e.g., Nass & Moon, 2000), one means of addressing these issues is the use of various simulated social approaches (e.g., persuasion, Lee & Liang, 2019; use of expressive gesture, Salem, Eyssel, Rohlfing, Kopp, & Joublin, 2013) to build user trust, especially within dynamic environments that pose challenges in ensuring flawless operation.

Schaefer, Chen, Szalma, and Hancock (2016) highlight trust as being essential in the formation and development of collaborative relationships, and key to supporting effective Human-Robot Interaction (HRI). However, the possibility of incorporating social behaviours into the design of synthetic social agents specifically to affect user perceptions raises important implementation and ethical issues in this still developing field (Baker, Phillips, Ullman, & Keebler, 2018). As the sophistication of interactive synthetic social agents increases, the relevance of social psychological models for both underpinning agent design and for understanding user responses will also increase (e.g. Dautenhahn, 2007; Fong, Nourbakhsh, & Dautenhahn, 2003a). However, quantifying a complex emotional response into variables that can be measured through controlled

experimentation is difficult, and meaning is inevitably lost. This can create an 'object conflict' (Hess, 2007), in which two parties believe that they hold the same definition or idea about a topic but actually do not. Thus, they may not realise that the reason they cannot come to an agreement is because they are not actually talking about the same thing. In a similar way, applying established approaches derived from earlier human-computer interaction studies to seemingly autonomous agents has revealed important epistemological issues as the field of HRI evolves (Baker et al., 2018); in particular, an object conflict in what it means to 'trust' a robot.

Most highly-cited or prominent papers describe models of user trust towards automation and autonomous agents in terms of capability and reliability in task performance (e.g., Hancock et al., 2011; Lee & See, 2004; Muir, 1994). Trust, therefore, is tacitly defined as an issue of whether the user feels confident that the machine can be relied upon to consistently provide the desired function, and will not break down or otherwise create a risk to users, property or task management. However, in the social cognitive models employed in human-to-human interaction, trust is also regarded as an element of warmth or likability, not solely of an agent's capability to carry out a task as expected (Fiske, Cuddy, & Glick, 2007; Rosenberg, Nelson, & Vivekananthan, 1968). This has as-yet unacknowledged influence over how users may tacitly be defining trust when engaging in HRI experiments, and how they may adapt social rules from human-human interactions to guide their interactions with robots, particularly in an interactive environment where the robot appears to be responding to social cues.

If one assumes that automation is a sophisticated tool to support users' progress towards goals, it may be intuitively appropriate to consider trust as a measure of the users' belief that the tool can capably support that effort. In these contexts, trust could be explored through the extent to which users monitor an automated system's activity and capacity to ensure reliability in task-performance (e.g., Ruff, Narayanan, & Draper, 2002). However, while trust-as-capability may suffice in the factory context, social questions acquire greater significance with reciprocity, when humans are engaged in dependent relations with synthetic agents Allen et al. (2011). Increasingly, advanced

social robots are expected to 'engage with' rather than just 'do', challenging the perspective that *relationships between people are qualitatively different from relationships between people and automation* (Lee & See, 2004) and in these contexts it may be more appropriate to consider trustworthiness in social-cognitive terms. Here, trust of others is identified as fundamental for interpersonal cooperation to be a success (Barczak, Lassk, & Mulki, 2010; McAllister, 1995). A synthetic agent may be considered extremely capable of reliably performing a set of expected actions, yet still inspire distrust, even rejection, because it is not felt to be trustworthy in social terms.

Collaborative working in HRI contexts (such as advanced manufacturing: *Factories of The Future*, European Commission, 2013; or assistive healthcare, Mukai et al., 2010) may also be contingent on trust between agents, particularly on the development of a person's ability to trust a non-human agent. Without a clear understanding of what 'trust' means to research participants, and how that might differ from how 'trust' is being framed by the researchers themselves, it is difficult to know exactly what we are measuring in HRI, or to compare experiments across the field. An understanding of the social components of trust is required to enable the ethical creation of robots that are meant to engender interactive trust in users, especially where robots are configured for behaviours designed to elicit a social response (such as asking questions or saying 'follow me'). Particularly with regard to trust in the standard testing of human-robot interactions with regard to error-handing procedures, designers and users may also want to ask:

- Does the robot's behaviour appear 'sincere' (Shim & Arkin, 2013) and in which circumstances is that meaningful (Risen & Gilovich, 2007)?

- Who determines which errors are suitably insignificant that they need not be designed out, and who holds responsibility for these errors should they occur (Sharkey, 2017)?

This paper discusses an experiment which uses social psychological concepts of trust to mitigate the negative effects of an autonomous tour-guide robot's errors to

begin to address that gap. Through this we contribute to a still-emerging theoretical understanding of trust in the context of presumed machine agency, and an exploration of nuances concerning the use of synthesized social behaviors when robots interact with humans. With these aims in mind, we first explore how trust has so far been approached in HRI research, then examine social definitions of trust arising from our data, and finally consider how this contributes to the creation of a general theoretical framework for exploring trust in social robotics which may better direct the field. By this we also hope to improve the ethical frameworks guiding the development of robot systems, which may be vital for both protecting users and shaping high-quality research.

## Theoretical Development

Trust has been an area of interest for collaborative working, personal assistance and social interaction in HRI for an extended period, and key papers in HRI and automation addressing trust scale development (e.g., Jian, Bisantz, & Drury, 2000; Muir, 1987) are highly cited and have directed the focus of this research. However, despite these enduring efforts to define, measure, and explore 'trust', it remains a concept which is not fully operationalised, potentially due to its meaning being tightly bound to the original context of reliability in which it was developed (see Cameron et al., 2015, for a discussion on this in terms of HRI).

In essence, it is not always clear in this research how measures of trust have been theorised, used and interpreted, and whether researchers are exploring 'trust' as a unified concept or as a series of factors. Even papers that clearly detail the process by which the trust scales have been developed display issues in their interpretation and use. For example, the 12-item survey in Jian et al. (2000) has been used both to score trust as a single metric (e.g., Kaniarasu, Steinfeld, Desai, & Yanco, 2012; Selkowitz, Lakhmani, Chen, & Boyce, 2015), and to more effectively represent two oblique factors: trust and distrust (Spain, Bustamante, & Bliss, 2008). Further to this, while Schaefer (2013) identifies that their developed scale captures elements of trust not measurable by Jian et al. (2000), they do not specify what these could be. Inspection of the scale

generation process in Schaefer (2013) suggests four candidate factors concerning trust that are ultimately reduced down to a single measure derived from 40- or 14-item scales. However, trust towards a robotic agent may be more multifaceted and complex than these current scales suggest; to this end, it may be useful to consider psychological models regarding trust as better describing the underlying values humans bring to these interactions.

A prominent model in the social-cognitive literature positions evaluations of others as being based on dimensions of competence and warmth (for a review, see Fiske et al., 2007), which account for substantial variance (over 80%, Wojciszke, Bazinska, & Jaworski, 1998) in people's response. Fiske et al. (2007) also highlight trust as primarily a construct of warmth, aligning it with concepts such as friendliness, helpfulness and morality, which are connected to *perceived intent*. These are contrasted to competency constructs (e.g., skill and intelligence) that are more related to *perceived ability*. Similar perspectives are discussed in the education literature concerning earning trust as teachers (Brookfield, 2015) and in the occupational psychology literature concerning teamwork and group formation (Cuddy, Glick, & Beninger, 2011). Finally, McAllister (1995) describes two forms of trust, cognitive and affective, namely the belief that another agent is 1) capable and 2) good-natured; Casciaro and Sousa-Lobo (2005) also emphasizes the importance of the affect/intent dimension of trust over perceived ability when selecting collaborative partners. These understandings indicate a substantive departure from how trust is usually framed in the HRI literature (e.g., Hancock et al., 2011), where it primarily indicates task-reliability of the robot, assessing competence rather than warmth, even when it is performing social tasks. Thus, the present challenge of accurately evaluating trust in HRI may in part arise because the available research methodologies, tools (Muir, 1987; Schaefer, 2013), and directions (Schaefer et al., 2016) are limited by the assumption that the user also applies 'trust' only to the robot's apparent ability and their responses are not influenced by its apparent intent.

Despite this, a more interpersonal perspective of trust is beginning to emerge in human-computer interaction (e.g., Kulms & Kopp, 2018) and in robotics (e.g., Cameron

et al., 2015; van Straten, Peter, Kühne, de Jong, & Barco, 2018), making a theoretically-grounded understanding of trust in robots as social agents timely. Particularly in studies using a self-reporting design, there may be unexamined object conflicts between 'trust' as conceived by the scale developers and 'trust' as understood by researchers implementing the scale, or between researchers and participants answering the questions. Additionally, as research shows that individuals can regard even text-based computer systems as social agents (Reeves & Nass, 1996) the anthropomorphic qualities of robots may further emphasize their social agency (Fong, Nourbakhsh, & Dautenhahn, 2003b), particularly where these have communicative functions, increasing the likelihood that people will consider *and trust* the robot from a social (warmth, rather than competence) perspective. We therefore suggest failure-of-competence experiments as a useful way of integrating social psychological concepts into the field.

RQ1 Can trust towards social robots be meaningfully represented through social cognitive perspectives?

**Making Errors**

Errors made during autonomous actions outside of strictly controlled environments are inevitable. These include, but are by no means limited to: drawing inferences from noisy sensor data, including sensor anomalies; mechanical and computational failure during operation; and/or failure to respond to unanticipated changes in the environment. Thus, it is imperative to understand user response to a system that has demonstrated issues in performance reliability, particularly as open environments may be sufficiently complex that it is not feasible for designers and programmers to prepare specific behaviors for anything but the most likely circumstances encountered. However, failure is also the area where psychological aspects, such as perceived warmth or intent in the response to the error, are most likely to influence user response (Honig & Oron-Gilad, 2018).

Typically, an HRI scenario entails an end goal for the user, the robot, or the two

as a collaborative team; robot errors that draw either agent away from their goals may have adverse effects on user experience which can be measured. In general, the literature suggests errors negatively impact user views; for example, in constructed scenarios comparing faulty and non-faulty versions of the same robot, users describe the faulty robot as less competent and have less favorable intentions to use it (Brooks, Begum, & Yanco, 2016). Errors in memory and reasoning or erratic behaviour may result in users regarding the robot as being less competent and less reliable (Ragni, Rudenko, Kuhnert, & Arras, 2016) or untrustworthy (Salem, Lakatos, Amirabdollahian, & Dautenhahn, 2015), while navigation errors in a mobile robot may discourage use of its autonomous systems (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013; Goodrich, Olsen, Crandall, & Palmer, 2001).

Because they adversely affect belief that the robot is useful or usable, in turn negatively affecting people's intention to use again (Venkatesh, 2000), it is obviously beneficial to prevent errors before their occurrence; to use feedback to iteratively design them out, as is standard practice in HCI systems (Norman, 2016); or to develop training which can support users both in accommodating errors and in maintaining use of the automated systems (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & Moray, 1992, 1994; Parasuraman & Riley, 1997). Advanced warning of the possibility of an error (Lee, Kiesler, Forlizzi, Srinivasa, & Rybski, 2010) or notification of a robot's confidence in a decision can support user trust towards the robot (Kaniarasu, Steinfeld, Desai, & Yanco, 2013) by supporting users to develop a clear mental model of a system's processes and capabilities (Norman, 2016), thus promoting usability. These approaches consume time, however, often require preparation on behalf of the user, and may not be available to users at the point of error. Some errors may only be apparent at the point of or after occurrence, or they may present their own difficulties in terms of processing power and speed of computation. Therefore, active recognition of robot errors during HRI, particularly with regard to users' understanding and response, has emerged as a valuable topic of research (e.g., Mirnig et al., 2017).

Much of this has concentrated on the context of the error, which may have a

significant impact on the response necessary from a robot (Correia, Guerra, Mascarenhas, Melo, & Paiva, 2018a). For example an occasional error from a robot designed for entertainment and play, such as Aibo (Lee, Peng, Jin, & Yan, 2006), might be interpreted by the user not as an error but as an element of the robot engaging in play, or a sign of its naivety and quirky 'personality', whereas errors that actually endanger humans will elicit a very different response (Adubor, St John, & Steinfeld, 2017). In other words, the context, task demands, and accuracy required in a robot's behavior may allow for some errors to occur in HRI that need not be designed away. While avoidance of possible errors will still remain integral to system development (as with all effective design in HCI, Norman, 2016), a potential paradigm shift in future design of effective robots (Salem et al., 2013) may instead consider the impact of a robot recognizing it has made an error and engaging in interaction to mitigate the social effects, or potentially even turning this into a net-positive for the interaction (e.g., Correia et al., 2018a; Hamacher, Bianchi-Berthouze, Pipe, & Eder, 2016; Kaniarasu et al., 2013; Lee, Kiesler, Forlizzi, Srinivasa, & Rybski, 2010). Central to this will be exploring the appropriate post-error actions and communication from the robot system to the user.

The current literature offers a small variety of scenarios in which people's responses to robotic errors and recovery are experimentally observed. Robot social recovery strategies such as offering an apology after an error (Hamacher et al., 2016; Lee, Kiesler, Forlizzi, Srinivasa, & Rybski, 2010; Robinette, Howard, & Wagner, 2015), justifying why the error occurred (Correia et al., 2018a), blaming itself for the error (Groom, Chen, Johnson, Kara, & Nass, 2010; Kaniarasu & Steinfeld, 2014), asking for assistance (Brooks et al., 2016), and promising to do better (Robinette et al., 2015) have all been explored. Participants' responses indicate the robots' social recovery strategies can mitigate robotic error's negative effects both in general (Brooks et al., 2016), and specifically on its capability (Lee, Kiesler, Forlizzi, Srinivasa, & Rybski, 2010), its likability (Groom et al., 2010; Kaniarasu & Steinfeld, 2014; Lee, Kiesler, Forlizzi, Srinivasa, & Rybski, 2010), and on people's intention to use (Hamacher et al.,

2016; Robinette et al., 2015). Of note, the apology strategy deployed in Kaniarasu and Steinfeld (2014) promoted likeability but lowered user perceptions of capability, as would be anticipated given the social cognitive model's oblique dimensions for these (Rosenberg et al., 1968). Social recovery strategies have also mitigated impact on trustworthiness across several measures: as a single reported score based on Schafer's (Schaefer, 2013) 14-item scale on the functionality of the robot (Correia et al., 2018a), as an ad-hoc single item measure on whether a user trusted the robot or not (Robinette et al., 2015), and as an ad-hoc single item score of how much a user trusts the robot (Hamacher et al., 2016).

Crucially, each of the above studies uses social means to address the errors and, of these, only two (promising to do better, justifying the error) specifically target the issue of capability in the robot's performance. An approach such as apologizing for the error (Hamacher et al., 2016; Robinette et al., 2015) is a behaviour more closely reflecting warmth-related constructs such as honesty and sincerity than capability, nonetheless, it still supported people's trust and intentions to use. As Casciaro and Sousa-Lobo (2005) suggest in the context of human-human affiliation, this personable dimension may create a human-machine affiliation response which has an important impact on trust of non-human social agents.

RQ2 Which social recovery strategies are more effective in supporting people's trust towards, and intention to use, an error-prone robot?

**Hypotheses**

We anticipate that, in line with the social cognitive model (Fiske et al., 2007), that following an error, a robot deploying social cognitive responses of either a *warmth* focused recovery strategy (apology) or a *competency* focused strategy (explanation) will have a positive impact on its *likability* and *capability* respectively. We further anticipate that an apology-based strategy will have a negative impact on perceptions of capability (following Kaniarasu & Steinfeld, 2014). Based on the social cognitive model, we anticipate that people may draw some distinctions between trust of a robot relating to

its warmth and competency in line with the above processes. Following Casciaro and Sousa-Lobo (2005), likability and warmth-related trust outcomes would have a greater positive impact on people's intention to use the robot than capability and competency-related trust outcomes. To explore our research questions regarding the potential contrasting influences of a robot's synthetic representations of warmth and capability on people's trust of, and intentions to use, a robot we consider the following hypotheses:

H1 Apologies for an error support people's perspectives of a robot's likability

H2 Explanations for an error support people's perspectives a robot's capability

H3 Apologies reduce people's perspectives of a robot's capability

H4 Trust towards the robot, represented in both warmth and competency terms, is affected in line with H1-3 respectively.

H5 Likability, and warmth-based trust, have greater effect than capability, and competency-based trust, on intention to use

## Method

Our experiment was designed to explore the potential of these social strategies of trust development in social robotics, and develop an empirical foundation for using trust models inspired from human-human interaction research. To achieve this, we constructed a demonstration of a mobile tour guide robot making navigational errors when leading a person to a target destination: a scenario designed to challenge perceptions of reliability. This allowed us to examine how people use multifaceted evaluations of trust through having the robot use varying social recovery strategies - *apologizing for the error* or *demonstrating knowledge of why the error occurred*. This was determined by user ratings of various trust-related words to describe the robot, which we predicted would align with factors resembling the social cognitive dimensions and demonstrate that these social recovery strategies had indeed had a positive impact.

## Design

A 2x2 independent measures design was implemented. The four conditions comprised videos of an HRI scenario with the presence or absence of key statements regarding i) the robot identifying how to correct its mistake, and ii) the robot apologizing for making a mistake. Participants were randomly allocated to a single condition and the Qualtrics survey engine prohibited participants repeating the survey.

## Participants

Participants were recruited through university staff and student volunteer mailing lists and through social media, and were offered a chance to win gift vouchers for their contribution to the study. In total, 549 people responded to the study invitation, however, 138 participants did not answer any questions in the surveys and so are excluded from the analysis. Attrition in online studies, such as this one, is expected (Hoerger, 2010), although the rate observed is higher than commonly seen. While we do not have data for participants' reasons they withdrew from the study early, the chance to win vouchers may have been sufficient to inflate population interest in taking the study but not to motivate full engagement (LaRose & Tsai, 2014).

Further participant exclusions were applied based on three criteria: 1) having already seen the HRI-scenario video and so a non-naive sample (32 excluded); 2) monitored play duration of the video being under 60 seconds and so not sufficient to reach the critical scene (26 excluded); and 3) no indication they had paid attention to the HRI video scenario in their descriptive response, e.g. "Very interesting" vs "A robot showing someone around a building" (27 excluded). Initial inter-rater agreement on participant written responses to include was 99.41% and 100% after discussion. After exclusions were applied, 326 participants were included for analysis. Sample size calculations for the study indicated that 326 participants would be sufficient to detect an effect size $F$ .15 in a 2x2 design at a two-sided significance level of .05 with 80% power.

User demographics of age, nationality and gender were collected, given that each

of these had previously been identified as impacting attitudes towards new technology and intentions to use robots (De Graaf & Allouch, 2013). Of the included participants, 169 identified as female, 139 as male, and 18 declined to answer or identified as an alternative gender. The large majority of participants, 211, were from the UK, while 97 were from overseas (38 from East Asia, 36 from Europe, 15 from North America, 4 from the Middle East, 3 from South America, and 1 from Africa) and 18 declined to provide this information. The mean age was 26.52 years (SD 9.97; with range from 18 to 62 years); sixteen participants declined to provide this information.

**Procedure**

Online distribution of pre-recorded HRI scenarios for participants' evaluation is a widely used approach (e.g., Kiesler, Powers, Fussell, & Torrey, 2008; Shen, Tennent, Claure, & Jung, 2018; Srinivasan & Takayama, 2016) that enables researchers to reach much more diverse audiences and in greater numbers than may be practical with 'live' HRI experiments. This approach is beneficial for exploratory work for its potential to identify small effects worthy of following up in live HRI experiments, and as a precaution against underpowered studies, which may produce difficult-to-replicate findings.

The study itself was delivered through the online survey engine Qualtrics. Participants were first presented with an information sheet on the study and a consent form for participation. On signing the consent and agreeing to participate, individuals were randomly allocated to view one of the four videos described below. The duration that each participant spent on the video page was silently tracked and used as an indicator of their engagement as participants who spent less than 60 seconds on the page would not have had the opportunity to view the critical interaction scene and were hence excluded from subsequent analysis. Participants then completed a brief statement about the video, again used as an index of their engagement. Participants who described either the robot or the scenario, even in broad terms, had their responses in the study included in the analysis.

Following the video, participants completed the Godspeed Questionnaire measures for perceived likability and capability (Bartneck, Kulić, Croft, & Zoghbi, 2009). They were then asked about intentions to use the robot and finally demographic details. At the end of the questionnaire participants could share contact details for a chance to win gift vouchers. The study took participants approximately 10 minutes to complete.

## Materials

**HRI Videos.** Four videos of user interaction with the robot were filmed ahead of distribution of the study. The videos represented a typical scenario of using ROBO-GUIDE: an individual encounters the mobile robot and is led to the room of their choosing. The location of the videos is the interior of a building likely familiar to the university staff and students invited to participate in the study, as it is based on campus and has many teaching facilities; this location was further selected due to its local notoriety as being difficult to successfully navigate.

ROBO-GUIDE is built using the Pioneer LX platform, manufactured by Adept MobileRobotics Law et al. (2015). The system is a wheeled platform (approx 50cm width, 70cm length, & 45cm height) that uses a laser scanner for indoor mapping and navigation; it also houses proximity and impact sensors and speech synthesis software. These elements make it a suitable candidate as a prospective assistance or guidance robot for the HRI scenario. While the robot is equipped to autonomously navigate the building, researchers pre-programmed its movements and responses through ROS for the purposes of the video. The Pioneer platform has previously been deployed in studies using navigation assistant scenarios (Scheggi, Aggravi, & Prattichizzo, 2016; Wang & Christensen, 2018).

The videos each followed the same general storyboard: a user is greeted by the robot and instructed to follow it to their target destination. The robot and user then travel along a corridor to an elevator and the user presses the appropriate elevator control as instructed by the robot. Robot error was presented as another person disembarking the elevator at a level differing from the user's destination; the robot and

the user also disembark at this point. This type of error is identified as a plausible outcome from the robot using noisy environmental cues in elevators to detect the current level (McAree et al., 2015). After a brief exploration of the incorrect level, the robot identifies that it must return to the elevator. Critically, at this point, the four videos differ:

i No further dialogue from the robot

ii The robot apologizes for the mistake (See Figure 1)

iii The robot identifies cues that it is on the incorrect level

iv The robot communicates both (ii) and (iii)

All videos then show the robot and user using the elevator to reach the correct level and arriving at the target destination. The four films have minimal differences as the same non-critical scenes were used for each and the critical scenes differed only in the audio of the robot's speech and subtitles. Each video lasted approximately 110 seconds.

**Measures.** Participants' views towards the robot were assessed using the Godspeed Questionnaire sections for *Capability* and *Likability* (Bartneck et al., 2009), presented in a random order for each participant. Godpseed was chosen for three reasons; first for its widespread use in HRI studies; second the two subscales used resemble that of the social cognitive framework (Fiske et al., 2007); and crucially, third, the scale was not developed to explore trust so conflicts concerning 'trust' as an ontological object are less likely to arise. For both measures of capability and likability, participants completed five items of five-point semantic differentials (e.g., for capability: Incompetent/Competent; for likability: Kind/Unkind). Both measures had a high reliability across items (Cronbach's $\alpha$ for capability = .81; for likability = .87).

Participants' descriptions of trust towards the robot were explored using terms drawn from the scale developed by Jian et al. (2000), which uses a grounded approach to identify terms strongly associated with trust in automation and clusters these to


*Figure 1*. ROBO-GUIDE apologizing to a user for arriving at the wrong level

form 12 items exploring people's beliefs about automated systems. Of specific interest to the current study are the terms associated with the social experience of trust: *capable, competent, confident, deceptive, false, honest, honorable, incapable, incompetent, loyal, misleading, reliable, trustworthy, unreliable, unsure, untrustworthy.* Participants scored the extent to which they would describe the robot as each of the terms from 1 (not at all) to 7 (extremely). Principal component analysis was used to identify construct dimensionality of the measures.

Participants' intentions towards using ROBO-GUIDE were assessed using a 4-item measure (Cronbach's $\alpha = .77$). Participants rated their agreement with 4 statements (e.g., I would use a ROBO-GUIDE in an unfamiliar place) on a 7-point scale ranging from *strongly agree* to *strongly disagree.* Items for this measure were presented in random order and included positive-intention and negative-intention statements. This measure has been developed and used in similar research concerning intentions to use mobile robots (Cameron, Loh, et al., 2016).

## Results

### Randomization Check

Preliminary testing of demographic data across conditions was conducted before analysis of the survey data. Participants were evenly distributed across conditions, with a minimum of 79 participants per cell ($\chi^2(1, N = 326) = .13$ $p = .66$).

There were no observed significant differences in age or gender between the 326 respondents that were included in the analysis and the 223 who were excluded (of which 74 completed demographic data). However, there was a significant difference between those included and excluded in terms of nationality ($\chi^2(1, N = 399) = 12.79$ $p < .01$). Regarding participants included in the study, UK nationality was represented approximately 2:1 compared to overseas representation. For respondents excluded from the study, UK:Overseas representation was approximately 1:1.

For each condition, there were even distributions of participants in terms of gender and nationality (Maximum $\chi^2(1, N = 326) = 4.99$ $p = .08$). There was a significant difference between conditions for age for the Competency condition only ($F(1,306) = 3.81$, $p = .05$); participants in the Competency conditions were slightly younger than those in the No-Competency conditions (M = 25.43, SD = 14.05; M = 27.63, SD = 14.05); this was a very small effect observed (Partial eta$^2$ = .01).

Participants' age shows a weak negative correlation with describing the robot as capable (r = -.13, n = 310, $p = .02$); age does not significantly correlate with any other measures used in the study. While we do not offer a theoretical account for this correlation, in the interests of transparency, results are presented with and without age as a covariate for measures of capability.

### Liking

There was a significant main effect of the robot stating an apology on participants' reports of liking the robot ($F(1,322) = 11.93$, $p < .01$), supporting Hypothesis 1. Specifically, participants who observed the robot offering an apology (across scenarios ii & iv: M = 3.67, SD = .74) reported liking the robot more than those who did not

(across scenarios i & iii: M = 3.38, SD = .74); this is a small effect (Partial eta$^2$ = .04).

There was no significant main effect of the robot stating its competency on participants' reports of liking the robot ($F(1,322) = 2.04$, $p = .15$). There was also no significant interaction effect observed between the robot's statements of competency and of apology on participants' reports of liking the robot ($F(1,322) = .60$, $p = .44$).

**Capability**

There was a significant main effect of the robot stating its competency on participants' reports of the robot's capability ($F(1,322) = 6.13$, $p = .01$), supporting Hypothesis 2. Those who observed the robot stating its competency (across scenarios iii & iv: M = 3.78, SD = .63) regarded the robot as more capable than those who did not (across scenarios i & ii: M = 3.60, SD = .64); this is a small effect (Partial eta$^2$ = .02).

There was a further significant main effect of the robot stating an apology on participants' reports of the robot's capability ($F(1,322) = 11.12$, $p < .01$), supporting Hypothesis 3. Those who observed the robot offering an apology (across scenarios ii & iv: M = 3.56, SD = .64;) regarded the robot as less capable than those who did not (across scenarios i & iii: M = 3.81, SD = .63), but this is a small effect (Partial eta$^2$ = .03). There was no significant interaction effect observed between the robot's statements of competence and of apology (scenario iv) on participants' reports of the robot's capability ($F(1,322) = 1.38$, $p = .24$).

As outlined in the Randomization Check, there was a significant negative correlation between participant age and perceptions of capability. Controlling for participant age did not materially affect the outcomes observed for either competency ($F(1,305) = 5.76$ $p$ =.02) or apology ($F(1,305) = 9.63$, $p < .01$) conditions. There was no significant interaction effect ($F(1,305) = 2.09$, $p = .15$).

**Trust**

The principal component analysis (PCA) was conducted using SPSS version 25. For the PCA, all 16 items concerning trust were included in a full data set analysis using a Direct Oblimin rotation with Kaiser Normalization (delta = 0), given the

oblique dimensions of warmth and ability in the social cognitive model (Fiske et al., 2007) and trust and distrust found in Jian et al.'s scale (Spain et al., 2008). Initial extraction of components was based on Eigenvalues of above 1.0 (see Table 1) and three components were extracted at this level; examination of the scree plot reflected this outcome well. The Kaiser-Meyer-Olkin measure of sampling adequacy (.89) and the $\chi^2$ for Bartlett's test of sphericity (1943.34) indicate that it is suitable to conduct a PCA on the items (Hair, Black, Babin, Anderson, & Tatham, 2006).

Table 1

*Total Variance with all items included*

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cum.% | Total | % of Variance | Cum.% |
| 1 | 5.88 | 36.77 | 36.77 | 5.88 | 36.77 | 36.77 |
| 2 | 2.43 | 15.20 | 51.97 | 2.43 | 15.20 | 51.97 |
| 3 | 1.14 | 7.12 | 59.10 | 1.14 | 7.12 | 59.10 |
| 4 | .88 | 5.55 | 64.65 | | | |
| 5 | .79 | 4.98 | 69.64 | | | |
| 6 | .64 | 4.01 | 73.65 | | | |
| 7 | .63 | 3.98 | 77.64 | | | |
| 8 | .51 | 3.23 | 80.88 | | | |
| 9 | .49 | 3.09 | 83.97 | | | |
| 10 | .44 | 2.80 | 86.77 | | | |
| 11 | .43 | 2.73 | 89.50 | | | |
| 12 | .38 | 2.43 | 91.94 | | | |
| 13 | .37 | 2.34 | 94.28 | | | |
| 14 | .33 | 2.10 | 96.38 | | | |
| 15 | .32 | 2.00 | 98.38 | | | |
| 16 | .25 | 1.61 | 100.00 | | | |

Items were considered to load on a component if the loading was greater than or equal to .50. In the analysis of rotated components, poorly loading items (scoring under a .50 benchmark) were removed and the PCA repeated. Through this iterative process, two items were discounted from the first component: confident (.49) and unsure (-.41) [1] and the final variation from extraction of components is shown in Table 2. Inspection of the three components and the items loaded points towards three latent constructs: Performance, Integrity, and Deceit. The Pattern matrix for item loadings is shown in Table 3 and indicates that all included items load with a minimum of .57.

Table 2

*Variance after low-loading items removed*

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cum.% | Total | % of Variance | Cum.% |
| 1 | 5.39 | 38.53 | 38.53 | 5.39 | 38.53 | 38.53 |
| 2 | 2.17 | 15.48 | 54.01 | 2.17 | 15.48 | 54.01 |
| 3 | 1.14 | 8.11 | 62.12 | 1.14 | 8.11 | 62.12 |
| 4 | .83 | 5.92 | 68.03 | | | |
| 5 | .64 | 4.59 | 72.62 | | | |
| 6 | .61 | 4.36 | 76.99 | | | |
| 7 | .53 | 3.81 | 80.80 | | | |
| 8 | .49 | 3.49 | 84.29 | | | |
| 9 | .44 | 3.15 | 87.44 | | | |
| 10 | .41 | 2.92 | 90.35 | | | |
| 11 | .40 | 2.87 | 93.22 | | | |
| 12 | .36 | 2.60 | 95.82 | | | |
| 13 | .33 | 2.32 | 98.14 | | | |
| 14 | .26 | 1.86 | 100.00 | | | |

[1] we acknowledge how close to threshold the first item is and how close to threshold the second would be at the lower threshold of .40 - use of the components with alternative thresholds does not substantively change trends observed across the experimental conditions

Table 3

*Factor Loadings (Items loading onto factors are shown in bold)*

| | Component | | |
|---|---|---|---|
| | Performance | Integrity | Deceit |
| Competent | **.88** | .11 | .16 |
| Reliable | **.81** | .11 | .05 |
| Capable | **.75** | .19 | .04 |
| Incapable | **-.64** | .17 | .21 |
| Incompetent | **-.64** | .21 | .29 |
| Unreliable | **-.63** | .12 | .24 |
| Trustworthy | **.57** | .48 | .07 |
| Loyal | .02 | **.83** | -.03 |
| Honorable | .00 | **.81** | .03 |
| Honest | -.09 | **.67** | -.25 |
| Deceptive | -.09 | -.04 | **.82** |
| False | -.03 | -.02 | **.72** |
| Untrustworthy | .10 | -.15 | **.70** |
| Misleading | .21 | .04 | **.66** |

Finally, the internal reliability of the components was evaluated through calculating Cronbach's alpha for each of the three components. The standard score for acceptable alpha is .7 (Hair et al., 2006) and the three components ranged from acceptable to high scores: Performance = .87, Integrity = .72, Deceit = .75

Participants' ratings of the robot for each of the three components were as follows: Performance M = 5.13 (SD = .96); Integrity M = 4.15 (SD = 1.36); Deceit M = 2.32 (SD= 1.03), indicating that participants rated the robot as at least moderately capable in performance and integrity but not particularly deceitful.

There were no significant main effects or interaction effects from conditions for

participants' ratings of the robot's integrity; this outcome do not support the first part of Hypothesis 4 - that results for integrity would align with those for Hypothesis 1. There was a significant main effect for the robot stating its competency on participants' ratings of deceit ($F(1,311) = 7.07$ , $p < .01$). Individuals who observed the robot highlight its competency rated the robot as showing less deceit than those who did not (M = 2.16, SD = .99; M = 2.47, SD = .99). This is a small effect (partial eta$^2$ = .02).

There was a significant main effect of the robot stating an apology on participants rating of their trust in the robot's performance ($F(1,311) = 10.27$, $p = .001$). Participants who observed the robot offer an apology (across scenarios ii & iv: M = 4.96, SD = .88) described the robot as being less capable in performance than those who did not (across scenarios i & iii: M = 5.30, SD = .88). This supports the final part of Hypothesis 4 - that these results would correspond with results for Hypothesis 3. This is a small effect (partial eta$^2$ = .03). There was no significant main effect for the robot stating its competence nor interaction effect on participants ratings in their trust of the robot's performance.

**Intentions**

There was a significant main effect of the robot stating an apology on participants' reported intentions to use the robot ($F(1,309) = 3.94$, $p < .05$). Participants who observed the robot offer an apology reported greater intention to use the robot than those who did not (M = 4.57, SD = 1.37; M = 4.24, SD = 1.37); this is a small effect (Partial eta$^2$ = .01).

There was no significant main effect of the robot stating its competence on participants' reported intentions to use the robot ($F(1,309) = 1.62$, $p = .39$). There was also no significant interaction effect observed between the robot's statements of competency and of apology on participants' reported intentions to use the robot ($F(1,309) = .01$, $p = .96$).

A multiple linear regression was calculated to explore participants' perceptions of the robot's characteristics on their intentions to use the robot. A stepwise approach

including the following measures as independent variables was used: Likablity, Capability, Trust(Performance), Trust(Integrity), and Trust(Deceit). Before the regression was conducted, the correlation matrix was inspected for high collinarity between variables, the strongest correlation was between Capability and Trust(Performance) where r = .65, scoring below the threshold of .70 for excessive collinearity; no variables were excluded based on collinearity ahead of conducting the regression.

The stepwise multiple linear regression identified two variables for inclusion in the final model: Likability and Trust(Integrity). Taken together these variables significantly predicted people's intention to use the guide ($F(2,310) = 81.71$, $p < .001$) and explained 34% of the variance in peoples intention to use the guide. Of the individual predictors, Likability had a higher standardised beta value (beta = .53, p < .001) than Trust(integrity) (beta = .13, $p = .01$). The remaining measures were not identified as being significant predictors (Capability $p = .36$; Trust(Performance) $p = .14$; Trust(Deceit) $p = .17$). Collectively these outcomes support Hypothesis 5 - that perceived likability and warmth-based trust (integrity) have a greater effect on intention to use than perceived capability and competency-based trust. Collinearity between the included variables was low (Variance Inflation Factor of 1.15) and no responses from participants were identified as being outlier data (i.e., all cases scored below the Mahalanobis Distance critical value of 13.82 for two independent variables).

## Discussion

The results indicate that brief, targeted interactions from a robot can significantly impact individuals' attitudes and intentions towards it. Specifically, a robot offering an apology for an error supports individuals' perceptions of its likability and in turn individuals' intentions to use the robot. In contrast, while a robot communicating its competence supports individuals' perceptions of its capability and a robot apologizing diminishes perceptions of its capability, neither appear to impact on individuals' intentions to use the robot.

The current findings offer support for the use of particular social approaches to mitigate potential adverse effects of robot error (RQ2). With regard to hypotheses offered: an apology offered by the robot after an error supported people's liking of the robot (H1); clear identification of how the robot detected the error supported people's perceptions of the robot's capability (H2); an apology offered by the robot after error limited people's perceptions of the robot as capable (H3); while people represented trust for the robot along social cognitive lines, these factors were only partially influenced by the apology and competency-based strategies: apologies lowered competency-based trust (H4); last, people's liking of the robot and their reports of warmth-based trust had greater positive impact than their perceptions of its capability and competency-based trust on their intentions to use (H5).

These findings conform to the broad model that people treat autonomous agents as having social qualities, and that people draw from human-human interaction experience in scenarios involving non-living agents (Nass & Moon, 2000). Specifically, the findings that a robot's use of warmth- or competency-based social cognitive strategies after an error support people's views of the robot along these dimensions align well with related studies in HRI (Honig & Oron-Gilad, 2018) and mirror findings in the wider literature on apologies in human-human interaction (Kim, Ferrin, Cooper, & Dirks, 2004) and people's interaction with non-human agents, such as companies (Xie & Peng, 2009). Of note, the particular finding that warmth-based strategies can impede perceptions of competency reflects outcomes in related literature for human-human interaction (Kim, Dirks, Cooper, & Ferrin, 2006) and HRI research (Kaniarasu & Steinfeld, 2014). As predicted in models of affiliation (Casciaro & Sousa-Lobo, 2005; Shazi, Gillespie, & Steen, 2015), liking and warmth had greater influence on people's intentions to use than capability and competency. Together, these outcomes suggest that people's perceptions of a robot's competence at a task may be sidelined by other social factors when it comes to engaging with social robots in assistive contexts.

It further appears that people evaluate the social robot's trustworthiness in similar social-cognitive terms (Fiske et al., 2007), as human-to-human interaction

(RQ1). Specifically, we identify three components relating to trust: Performance, Integrity, and Deceit. The differing perceptions for these across conditions and relative impacts on participants' intention to use the robot suggest participants have evaluated the robot across social cognitive dimensions and that the PCA outcomes are not simply a product of semantically grouping similar terms.

While it was not anticipated the PCA would reveal the third component, *deceit*, this does align with HRI literature (Spain et al., 2008). Terms used in this study to explore trust were drawn from scale development in Jian et al. (2000); their final scale has been indicated to examine two oblique factors of trust and distrust (Spain et al., 2008). Arguments for distrust as being distinct from trust are also made for human-human interaction (e.g., Lewicki & Wiethoff, 2000), which may be worth considering further in HRI contexts.

Looking further at the factors for trust, the term *trustworthy* is of note for two related reasons. First, while it loads onto the first component *performance* it also has a high cross-loading onto *integrity* (close to the .50 threshold) and is the only item in Table 3 for this to occur. Secondly, in a procedurally similar study (Cameron, Collins, et al., 2016), the item *trustworthy* loads on to *integrity* (.69) with no cross-loading onto *performance* (.19). The different interaction contexts across the studies may offer explanation: in the current study, a mistake may prime individuals to consider trustworthiness in terms of the (error-prone) robot's reliability; whereas, outside an error context (i.e., Cameron, Collins, et al., 2016), reliable performance could be much less pertinent so trust perceptions may more concern a social robot's integrity. Together, these issues may point towards a term with high potential for object conflict (Hess, 2007), as it appears that some individuals group trustworthiness in robots with the sense of its Performance, while others group it with the sense of Integrity, and existing measures do not allow researchers to determine which.

This points to an important lacuna in HRI research, in that 'trust' is generally understood within this literature as corresponding to performance-based criteria (e.g., Hancock et al., 2011), rather than social criteria such as integrity, and therefore most

empirical research continues to be situated within that framing (Baker et al., 2018). In experiments concerning reliability and error during human-robot interaction, the emphasis on the error context may therefore encourage individuals to consider trust towards a robot more strongly in technical terms, and subliminate their reaction to it as a social agent even as this tacitly informs their response. As integrity-related criteria are in general not part of the tools used for measurement of HRI, their significance may be missed, and the understanding even of performance-related criteria may be thus diminished.

**Limitations**

The results presented here should be interpreted with three limitations in mind. First, the study was conducted as an online questionnaire in which participants are responding to video footage of another person's experience with HRI and addressing a hypothetical future interaction which may never take place. Current findings point towards suitable areas for live-study HRI experiments, but these may not necessarily yield the same positive outcomes (although it should be noted that live-HRI could also give stronger effects). The large sample size available from the online survey does, however, serve as a protective factor against finding false positives.

Second, and related to the first, the outcomes found across the study each are small effect sizes. Nonetheless, it is worth noting that the current effects found do come from marginally different conditions: the simple communication of one or two brief sentences. Potentially, these effects could be strengthened with greater or repeated emphasis made when communicating the themes of capability or warmth. Alternatively, and as mentioned previously, a live-study HRI experiment may indicate stronger effects owing to people experiencing the interaction rather than imagining it.

Last, the positive effects on capability do not contribute to intentions to use, in contrast with our expectations. This outcome may occur because the scenario is very 'low-stakes' compared to other HRI studies, and the robot's capability in navigating may be relatively unimportant to a user. Introducing a sense of real inconvenience (e.g.,

Brooks et al., 2016) or potential peril (e.g., Robinette et al., 2015; Robinette, Wagner, & Howard, 2013) may increase the salience of a robot's capability, which may in turn have an impact on people's intentions to use. However, it should be noted that recovery of trust may not be possible under sufficiently severe consequences (Correia, Guerra, Mascarenhas, Melo, & Paiva, 2018b). The findings presented here provide an opportunity for a further comparative study into circumstances where social based strategies are applicable.

**Future Directions & Applications**

The outcomes of the study point towards a potentially fruitful area to explore regarding trust in social and collaborative robotics. Where various contexts face radical change due to the emergence of social robotics, such as healthcare(Robinson, MacDonald, & Broadbent, 2014), education (Mubin, Stevens, Shahid, Al Mahmud, & Dong, 2013), and wider workplaces or the home (Royakkers & van Est, 2015; Vänni & Korpela, 2015), issues of acceptance and trust towards robots may persist (e.g., Hancock et al., 2011; Savela et al., 2018). An understanding that robots in these contexts are not just able to perform their roles but that these are done with *specific intention for the benefit of the user* may support in these issues, particularly where trust in others' capabilities *and intentions* is key, such as health (Cameron, Sarda Gou, & Sbaffi, 2020), and education (Brookfield, 2015).

There are many factors that are considered to affect user experience and trust during HRI (Hancock et al., 2011) - the specific behaviours from a robot during interaction being just one of them. It remains to be considered how the particular approaches used in this study are situated amongst these diverse competing or complementary factors and the comparative impact they may have. Future studies could focus on the use of multiple strategies from a social perspective to engender trust and acceptance. This could be done to identify most useful approaches to take or to compare the influence of these across different morphologies and interaction contexts. In terms of wider theoretical development, the current research may present a useful

starting point for experimentally exploring the emerging perspectives of trust in HRI as being social and multifaceted (e.g., Ullman & Malle, 2019) and its further integration into models of user acceptance and intentions to use robots (e.g., Razin & Feigh, 2020).

Practical implications may include designers and programmers further considering the use of social-based strategies to support HRI in collaborative contexts and understanding the 'rules' that people import during interaction (e.g., Nass & Moon, 2000). In navigation assistance contexts this may support with rapid uptake in urgent circumstances, such as building evacuation (Robinette, Howard, & Wagner, 2017). Clear indication of a robot's apparent intention may help users make informed choices in how to engage with the robot, if at all, and especially in circumstances where errors occur (e.g., Salem et al., 2015). Interactions could be further explored with consideration that users are likely to go through trial and error with robots (e.g., Hamacher et al., 2016), correct errors together, and potentially build the user experiences together. Examination of HRI scenarios as collaborative experiences may further highlight relevant social elements that reflect human-human collaborations, such as affective trust (Webber, 2008), that may not be revealed using current measures and approaches towards robot trust.

It is important for the authors to note that we do not view the work we have described here to be scale development per se. Such a venture will require the same thoroughness observed in dedicated scale-development papers that form current landmarks in HRI trust research (Jian et al., 2000; Muir, 1987; Schaefer, 2013), and is beyond the scope of what we set out to achieve. Rather, we hope this paper serves to direct further research towards a largely-overlooked aspect of understanding trust in robotics, particularly in the consideration of robots which will primarily function as synthetic social agents.

## Conclusions

This paper presents novel work addressing the relative benefits that simple communication from a mobile robot can have on user attitudes after a robot-caused

navigational error. In line with HRI literature, offering explanations for its error supports people's perceptions of the robot as being capable (Correia et al., 2018a), whereas offering apologies lowers this perception (Kaniarasu & Steinfeld, 2014). Conversely, a robot's apologies for an error support its likability (Lee, Kiesler, Forlizzi, Srinivasa, & Rybski, 2010). These outcomes, in conjunction with our novel findings that perceptions of likability and integrity but not capability positively affects intention to use, reflect established social cognitive theory Casciaro and Sousa-Lobo (2005); Fiske et al. (2007). These further indicate the importance of social aspects for trusting and using autonomous robots and, more generally, social cognitive models have an appropriate role to play in exploring HRI.

While the perspective that people have social expectations for synthetic social agents is historically well-documented (e.g., Foner, 1997; Reeves & Nass, 1996), the current work further expands the perspective that use of socially empathic behaviours by embodied synthetic agents may facilitate better user experiences in spite of imperfection. Socially appropriate recovery strategies deployed by the robot after an error could thus be a viable means for robots to address errors which they cannot invisibly self-correct. An attention to warmth in the programming of socially interactive robots could build user confidence in their operation, and support their earlier and wider-spread adoption.

References

Adubor, O., St John, R., & Steinfeld, A. (2017). Personal safety is more important than cost of damage during robot failure. In *Proceedings of the companion of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 403–403).

Allen, C., Wallach, W., Hughes, J. J., Bringsjord, S., Taylor, J., Sharkey, N., ... others (2011). *Robot ethics: the ethical and social implications of robotics*.

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 1–30.

Barczak, G., Lassk, F., & Mulki, J. (2010). Antecedents of team creativity: An examination of team emotional intelligence, team trust and collaborative culture. *Creativity and innovation management*, *19*(4), 332–345.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, *1*(1), 71–81.

Brookfield, S. D. (2015). *The skillful teacher: On technique, trust, and responsiveness in the classroom*. John Wiley & Sons.

Brooks, D. J., Begum, M., & Yanco, H. A. (2016). Analysis of reactions towards failures and recovery strategies for autonomous robots. In *25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 487–492).

Cameron, D., Aitken, J. M., Collins, E. C., Boorman, L., Chua, A., Fernando, S., ... Law, J. (2015). Framing factors: The importance of context and the individual in understanding trust in human-robot interaction. In *International conference on intelligent robots and systems (IROS), workshop on designing and evaluating social robots for public settings.*

Cameron, D., Collins, E., Cheung, H., Chua, A., Aitken, J. M., & Law, J. (2016). Don't worry, we'll get there: Developing robot personalities to maintain user interaction

after robot error. In *Conference on biomimetic and biohybrid systems* (pp. 409–412).

Cameron, D., Loh, E. J., Chua, A., Collins, E., Aitken, J. M., & Law, J. (2016). *Robot-stated limitations but not intentions promote user assistance.* Paper Presented at New Frontiers in Human Robot Interaction, 52nd Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB'16), Sheffield, UK.

Cameron, D., Sarda Gou, M., & Sbaffi, L. (2020). Trust in robot-mediated health information. In *The 29th ieee international conference on robot & human interactive communication (ro-man), workshop for trust, acceptance and social cues in human-robot interaction - scrita.*

Casciaro, T., & Sousa-Lobo, M. (2005). Competent jerks, lovable fools, and the formation of social networks. *Harvard business review*, *83*(6), 92–99.

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018a). Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 507–513).

Correia, F., Guerra, C., Mascarenhas, S., Melo, F. S., & Paiva, A. (2018b). Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 507–513).

Cuddy, A. J., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in organizational behavior*, *31*, 73–98.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, *362*(1480), 679–704.

De Graaf, M. M., & Allouch, S. B. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems*, *61*(12),

1476–1486.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction* (pp. 251–258).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697–718.

European Commission. (2013). *Factories of the future multi-annual roadmap for the contractual ppp under horizon 2020*. Luxembourg: Publications Office of the European Union.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77–83.

Foner, L. N. (1997). Entertaining agents: A sociological case study. *Proceedings of the First International Conference on Autonomous Agents*, *5*(12), 122-129.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003a). A survey of socially interactive robots. *Robotics and autonomous systems*, *42*(3-4), 143–166.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003b). A survey of socially interactive robots. *Robotics and autonomous systems*, *42*(3-4), 143–166.

Goodrich, M. A., Olsen, D. R., Crandall, J. W., & Palmer, T. J. (2001). Experiments in adjustable autonomy. In *Proceedings of ijcai workshop on autonomy, delegation and control: interacting with intelligent agents* (pp. 1624–1629).

Groom, V., Chen, J., Johnson, T., Kara, F. A., & Nass, C. (2010). Critic, compatriot, or chump?: Responses to robot blame attribution. In *Proceedings of the 5th acm/ieee international conference on human-robot interaction* (pp. 211–218).

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis . uppersaddle river.* NJ: Pearson Prentice Hall.

Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016). Believing in bert: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction. In *Robot and human*

*interactive communication (ro-man), 2016 25th ieee international symposium on* (pp. 493–500).

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

Hess, D. (2007). What is a clean bus? object conflicts in the greening of urban transit. *Sustainability: Science, Practice and Policy*, *3*(1), 45–58.

Hoerger, M. (2010). Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking*, *13*(6), 697–700.

Honig, S., & Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, *9*, 861.

IFR Statistical Department. (2019). *World robotics service robots 2019: Statistics, market analysis, forecasts, case studies.* VDMA Services GmbH.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Kaniarasu, P., Steinfeld, A., Desai, M., & Yanco, H. (2012). Potential measures for detecting trust changes. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction* (pp. 241–242).

Kaniarasu, P., Steinfeld, A., Desai, M., & Yanco, H. (2013). Robot confidence and trust alignment. In *Human-robot interaction (hri), 2013 8th acm/ieee international conference on* (pp. 155–156).

Kaniarasu, P., & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. In *Robot and human interactive communication, 2014 ro-man: The 23rd ieee international symposium on* (pp. 850–855).

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic

interactions with a robot and robot–like agent. *Social Cognition*, *26*(2), 169–181.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational behavior and human decision processes*, *99*(1), 49–65.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology*, *89*(1), 104.

Kulms, P., & Kopp, S. (2018). A social cognition perspective on human–computer trust: The effect of perceived warmth and competence on trust in decision-making with computers. *Frontiers in Digital Humanities*, *5*, 14.

LaRose, R., & Tsai, H.-y. S. (2014). Completion rates and non-response error in online surveys: Comparing sweepstakes and pre-paid cash incentives in studies of online behavior. *Computers in Human Behavior*, *34*, 110–119.

Law, J., Aitken, J. M., Boorman, L., Cameron, D., Chua, A., Collins, E. C., ... McAree, O. (2015). Robo-guide: Towards safe, reliable, trustworthy, and natural behaviours in robotic assistants. In *Towards autonomous robotic systems (TAROS) 2015* (Vol. 9287, p. 149-154).

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153–184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Lee, K. M., Peng, W., Jin, S.-A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, *56*(4), 754–772.

Lee, M. K., Kiesler, S., & Forlizzi, J. (2010). Receptionist or information kiosk: how do people talk with a robot? In *Proceedings of the 2010 acm conference on computer*

*supported cooperative work* (pp. 31–40).

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *Human-robot interaction (hri), 2010 5th acm/ieee international conference on* (pp. 203–210).

Lee, S. A., & Liang, Y. J. (2019). Robotic foot-in-the-door: Using sequential-request persuasive strategies in human-robot interaction. *Computers in Human Behavior*, *90*, 351–356.

Lewicki, R. J., & Wiethoff, C. (2000). Trust, trust development, and trust repair. *The handbook of conflict resolution: Theory and practice*, *1*(1), 86–107.

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal*, *38*(1), 24–59.

McAree, O., Aitken, J. M., Boorman, L., Cameron, D., Chua, A., Collins, E. C., ... Martinez-Hernandez, U. (2015). Floor determination in the operation of a lift by a mobile guide robot. In *European conference on mobile robots (ECMR)* (pp. 1–6).

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, *4*, 21.

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., & Dong, J.-J. (2013). A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, *1*(209-0015), 13.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Mukai, T., Hirano, S., Nakashima, H., Kato, Y., Sakaida, Y., Guo, S., & Hosoe, S. (2010). Development of a nursing-care assistant robot riba that can lift a human in its arms. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 5996–6001).

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81–103.

Norman, D. (2016). *The design of everyday things - revised and expanded edition.* Basic Books, New York.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230–253.

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016). Errare humanum est: Erroneous robots in human-robot interaction. In *Robot and human interactive communication (ro-man), 2016 25th ieee international symposium on* (pp. 501–506).

Razin, Y., & Feigh, K. (2020). The measure of trust between man and machine: A meta-analysis of trust metrics in hri. In *The 29th ieee international conference on robot & human interactive communication (ro-man), workshop for trust, acceptance and social cues in human-robot interaction - scrita.*

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

Risen, J. L., & Gilovich, T. (2007). Target and observer differences in the acceptance of questionable apologies. *Journal of personality and social psychology*, *92*(3), 418.

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. In *International conference on social robotics* (pp. 574–583).

Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, *47*(4), 425–436.

Robinette, P., Wagner, A. R., & Howard, A. M. (2013). Building and maintaining trust between humans and guidance robots in an emergency..

Robinson, H., MacDonald, B., & Broadbent, E. (2014). The role of healthcare robots for older people at home: A review. *International Journal of Social Robotics*, *6*(4), 575–591.

Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach

to the structure of personality impressions. *Journal of personality and social psychology*, *9*(4), 283.

Royakkers, L., & van Est, R. (2015). A literature review on new robotics: automation from love to war. *International journal of social robotics*, *7*(5), 549–570.

Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence: Teleoperators & Virtual Environments*, *11*(4), 335–351.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, *5*(3), 313-323.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). 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). ACM.

Savela, N., Turja, T., & Oksanen, A. (2018). Social acceptance of robots in different occupational fields: A systematic literature review. *International Journal of Social Robotics*, *10*(4), 493–502.

Schaefer, K. (2013). The perception and measurement of human-robot trust. *PhD Thesis, University of Central Florida*.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, *58*(3), 377–400.

Scheggi, S., Aggravi, M., & Prattichizzo, D. (2016). Cooperative navigation for mixed human–robot teams using haptic feedback. *IEEE Transactions on Human-Machine Systems*, *47*(4), 462–473.

Selkowitz, A., Lakhmani, S., Chen, J. Y., & Boyce, M. (2015). The effects of agent transparency on human interaction with an autonomous robotic agent. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 59,

pp. 806–810).

Sharkey, A. (2017). Can robots be responsible moral agents? and why should we care? *Connection Science*, *29*(3), 210–216.

Shazi, R., Gillespie, N., & Steen, J. (2015). Trust as a predictor of innovation network ties in project teams. *International Journal of Project Management*, *33*(1), 81–91.

Shen, S., Tennent, H., Claure, H., & Jung, M. (2018). My telepresence, my culture?: An intercultural investigation of telepresence robot operators' interpersonal distance behaviors. In *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 51).

Shim, J., & Arkin, R. C. (2013). A taxonomy of robot deception and its benefits in hri. In *IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 2328–2335).

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 52, pp. 1335–1339).

Srinivasan, V., & Takayama, L. (2016). Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4945–4955).

Ullman, D., & Malle, B. F. (2019). Measuring gains and losses in human-robot trust: evidence for differentiable components of trust. In *2019 14th acm/ieee international conference on human-robot interaction (hri)* (pp. 618–619).

Vänni, K. J., & Korpela, A. K. (2015). Role of social robotics in supporting employees and advancing productivity. In *International conference on social robotics* (pp. 674–683).

van Straten, C. L., Peter, J., Kühne, R., de Jong, C., & Barco, A. (2018). Technological and interpersonal trust in child-robot interaction: An exploratory study. In *Proceedings of the 6th international conference on human-agent interaction* (pp. 253–259).

Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control,

intrinsic motivation, and emotion into the technology acceptance model. *Information systems research*, *11*(4), 342–365.

Wang, S., & Christensen, H. I. (2018). Tritonbot: First lessons learned from deployment of a long-term autonomy tour guide robot. In *27th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 158–165).

Webber, S. S. (2008). Development of cognitive and affective trust in teams: A longitudinal study. *Small group research*, *39*(6), 746–769.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263.

Xie, Y., & Peng, S. (2009). How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology & Marketing*, *26*(7), 572–589.