On the Impact of Different Types of Errors on Trust in Human-Robot Interaction: Are laboratory-based HRI experiments trustworthy?

Rebecca Flook
Dpt. of Computer Science
University of Bristol
Bristol, UK
rebecca@flook.name

Anas Shrinah
Dpt. of Computer Science
University of Bristol
Bristol, UK
anas.shrinah@bristol.ac.uk

Luc Wijnen
Dpt. of Artificial Intelligence
Radboud University
Nijmegen, The Netherlands
Luc2.Wijnen@live.uwe.ac.uk

Kerstin Eder
Dpt. of Computer Science
University of Bristol
Bristol, UK
kerstin.eder@bristol.ac.uk

Chris Melhuish
Bristol Robotics Laboratory
Univ. of the West of England
Bristol Robotics Laboratory
Bristol, UK
chris.melhuish@brl.ac.uk

Séverin Lemaignan
Bristol Robotics Laboratory
Univ. of the West of England
Bristol Robotics Laboratory
Bristol, UK
severin.lemaignan@brl.ac.uk

Abstract

Trust is a key dimension of human-robot interaction (HRI), and has often been studied in the HRI community. A common challenge arises from the difficulty of assessing trust levels in ecologically invalid environments: we present in this paper two independent laboratory studies, totalling 160 participants, where we investigate the impact of different types of errors on resulting trust, using both behavioural and subjective measures of trust. While we found a (weak) general effect of errors on reported and observed level of trust, no significant differences between the type of errors were found in either of our studies. We discuss this negative result in light of our experimental protocols, and argue for the community to move towards alternative methodologies to assess trust.

Keywords: human-robot interaction, trust

Biographies

Rebecca Flook has just achieved a MSc in Computer Science with Distinction at the University of Bristol, UK. Her previous experience includes

a First-class honours in Adult Nursing and experience in the field of Speech and Language Therapy.

Anas Shrinah is a PhD student at the University of Bristol and member of the Trustworthy Systems Laboratory. He has a First-class honours BEng in Computer and Automation Engineering and MSc with Distinction in Robotics. Anas' research is focused on verification techniques for planning-based automated systems.

Luc Wijnen is currently an MSc student at the Bristol Robotics Laboratory, UK. He did his BSc in Artificial Intelligence at the Radboud University in Nijmegen followed by an MSc in Artificial Intelligence with a specialization in Robot Cognition at the same university.

Kerstin Eder is Professor of Computer Science and leads the Trustworthy Systems Laboratory at the University of Bristol. Much of her research is focused on specification, verification and analysis techniques to verify or explore a system's behaviour in terms of functional correctness, safety, performance and energy efficiency.

Chris Melhuish, BSc, MSc, PhD, CEng, FIET, FBCS is the founding member and Director of the Bristol Robotics Laboratory (BRL), a collaboration between the University of the West of England and the University of Bristol, and home to the UK's largest academic centre for multi-disciplinary robotics research. Chris holds professorial chairs at both Universities. His research interests include safe human-robot interaction, energetically autonomous robots, haptics and swarm systems. He has published over 250 peer reviewed papers and led numerous UK and EU research projects.

Séverin Lemaignan is Senior Research Fellow at the Bristol Robotics Laboratory, University of the West of England. He was awarded his PhD in Artificial Intelligence for Human-Robot Interaction in 2012, and since then, he has been focusing his research on social robotics, artificial theory of mind, and the human factors of HRI.

1 Introduction

As the demand for robotic co-workers increases, so does the need for trustworthy machines. Trust is a multi-faceted belief that is difficult to gain and easy to lose. One of these facets relates to the ability of a robotic assistant to carry out a prescribed task (B. Muir & Moray, 1996). Robots do not, as of yet, perform flawlessly, and as such investigating the effect of robot errors on the resulting trust levels is a well researched topic in human-robot interaction (HRI) (Hancock et al., 2011; Mirnig et al., 2017).

Salem, Lakatos, Amirabdollahian, and Dautenhahn (2015) suggest that there is a lack of adequate definitions of *trust*, specifically within HRI. They suggest that looking at definitions from similar fields, namely automation and human-computer interaction may assist in providing definitions, despite

the fact that these areas differ in terms of variety of interactions. Robots have indeed a greater, more human-like, physical manifestation that may result in varying levels of trust. Salem et al. conclude their investigation by noting that most definitions of trust in HRI pertain to concepts relating to reliability and predictability.

Moray and Inagaki (1999) define trust in automation as "an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of the designer's intentions, be relied on to achieve the design goals". B. M. Muir (1994) aimed to model the concept of trust by combining Barber (1983) and other research relating to human-machine trust. Their first model of human expectation of trust with robots includes in particular the ideas of "persistence, technical competency and fiduciary responsibility". J. D. Lee and See (2004) combine these expectations into three dimensions of trust: *purpose*, *process* and *performance*. Mayer, Davis, and Schoorman (1995) also define trust to have the following characteristics: ability ("the trustee competence in performing expected actions"); benevolence ("the trustee intrinsic and positive intentions towards the trustor") and integrity ("the trustee's adherence to a set of principles that are acceptable to the trustor"). In the rest of this article, we adopt the general definition by Moray and Inagaki: trust, in our context, is understood in term of the reliable realisation of expectations.

The notion of a right level of trust is discussed through existing literature. Hamacher, Bianchi-Berthouze, Pipe, and Eder (2016) state that some human-like behaviours lead to increased levels of trust, but might also have negative impacts when the "behaviour is deemed to cross a line". This is supported by Hancock et al. (2011) who describe that there are lower rates of satisfaction when interacting with robots that instil disproportionate trust levels in their human partner.

Research on the impact of errors is characterized by varying findings; ranging from the occurrence of errors making the robot seem more human-like, to resulting in a negative impact on trust (Salem, Eyssel, Rohlfing, Kopp, & Joulbin, 2013; Salem et al., 2015; Desai et al., 2012). Our aim is to further clarify these findings by providing new evidence on the effect of errors made by robotic co-workers, with the aim to understand the way in which robotic co-workers should be programmed, in direct relation to efficiency.

We present hereafter two independent studies that both investigate, not only the impact of errors on a participant's perceived level of trust in a robotic co-worker, but the effect of different types of error (*technical failures* versus *decision-level failures*) and the possible impact of the robot recognising and acknowledging these errors.

We are measuring trust using both subjective metrics (questionnaires) and behavioural metrics (based on proxemics), on two different robotic platforms (Aldebaran's Pepper and PAL's TIAGo).

Figure 1: Experimental setup for Study 1. Participants are sat in front of the robot; the wizard is sitting behind the participants, out of their field of view. The robot guides the assembly of a toy car by the participant, using the parts displayed on the table.

1.1 Factors Affecting Trust

To better identify how trust is affected in human-robot interaction, the factors that impact upon trust, both positively and negatively, need to be researched. These have been separated into three main areas, namely: robot, human and environmental factors and further subsections within each of these domains. This attempts to assess factors that are not just presented on the robot's behalf, whilst uncovering areas that need consideration.

1.1.1 Robot Factors

Robot Errors The most prominent robotic factor in relation to this research is robots making errors. Existing research reaches varying conclusions on the impact of these errors on trust. Corritore, Kracher, and Wiedenbeck (2003) report a greater negative impact on trust if multiple, less severe errors were made in comparison with one more severe error. Reiterated in later research, errors negatively impact upon perceived trustworthiness and reliability but do not however, affect the participants willingness to cooperate.

The presence of errors has been reported to result in increased anthropomorphism and likeability, despite a reduced task performance (Salem et al., 2013). Mirnig et al. (2017) found no significant impact of errors on a final perceived level of trust in a robotic assistant but also found an increase in

likeability. Guznov, Lyons, Nelson, and Woolley (2016) also found no statistically significant impact on self-reported trust levels in automation despite manipulating both error type and severity.

Although the research has shown that the presence of errors in HRI may have varying effects, one constant is found throughout existing literature, these errors can be compensated for. It is reported that participants appreciate a robot's attempt to apologize or rectify a situation where it had made an error (M. Lee, Kiesler, & Forlizzi, 2010). Whilst others conclude the perceived intelligence of the robot increased after having made a mistake and attempting to put it right, but only when the new method was error-free (Lemaignan, Fink, & Dillenbourg, 2014; Hamacher et al., 2016).

Mirnig et al. (2017) allowed for the classification of real errors into types; social norm and technical. They also highlighted that all robotic errors could be classed as technical from a roboticists point of view in contrast with a naive participant. The study defines the errors in the following ways; "a social norm violation (SNV) means that the robot's actions deviate from the underlying social script" and "a technical failure (TF) means the robot experiences a technical disruption that is perceived as such by the user".

Etiquette Parasuraman and Miller (2004) defined etiquette as "the set of prescribed and proscribing behaviours that permits meaning and intent to be ascribed to actions". They also studied the effect of etiquette on users' reported level of trust in an automated system.

Communication Style Studies have been carried out that attempt to analyse the preferred mode of communication in HRI; finding a robot with a more expressive interface that completed the task slower was more desirable than a highly effective robotic assistant that resulted in the participants reporting "feeling rushed" (Hamacher et al., 2016). Dautenhahn et al. (2005) reported 71 percent of participants would prefer a "human-like manner" of communication in a robot; including speech (Ray, Mondada, & Siegwart, 2008; Iwamura, Shiomi, Kanda, Ishiguro, & Hagita, 2011) and facial expressions (Sidner, Lee, & Lesh, 2003), specifically when they appear happy (Thrun, Schulte, & Rosenburg, 2000). Humans respond well to all forms of non-verbal communication attempts (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005), looking at the user (Bickmore et al., 2008) and referring to the user by name (Shiomi, Kanda, Ishiguro, & Hagita, 2006).

Behaviour transparency Transparency of a robot's behaviours can alter the amount of trust a human participant will instil in a robotic assistant. Wortham, Theodorou, and Bryson (2016) found that artificial agents that do not appear to have any other purpose other than to provide companionship seem unworthy due to a lack of no self-serving agency. Under the guise of

interacting in an assembly task this should result in the participant building some form of trust relationship in the robot, giving the experimenters a factor to measure.

1.1.2 Human Factors

Human Perceptions of Robots The Uncanny Valley (Mori, 1970) concept frames most of the research on trust in relation to robot appearance; presenting that humans find human-looking robots unnerving. Ray et al. (2008) highlighted facets of people's perceptions of robots; namely what they believe robots should look like. People responded they would prefer a robot to look like a small machine as opposed to resembling a living-being, such as a human, animal or other unspecified creature.

Dautenhahn et al. (2005) found that 40 percent of 28 people favoured the idea of robot companionship, but solely in relation to performing household tasks, in opposition to child and animal care or a personal relationship.

Previous Experience of Robots Bartneck, Suzuki, Kanda, and Nomura (2007) reported previous experience of robots could lead to less anxiety toward robots.

Personality Traits Nass and Lee (2000) reported that participants showed a preference towards robots exhibiting a personality type similar to their own, namely introverted or extroverted. Goetz, Kiesler, and Powers (2003) found that for personality traits such as seriousness and playfulness, people showed higher levels of cooperation with a robot displaying personality traits matching their own. Salem et al. (2015) found that participants that rated themselves as more extroverted and emotionally stable had higher levels of "psychological closeness" and "anthropomorphism" towards the robot and a more positive impression of the robot.

1.1.3 Environmental Factors

Severity of Human-Robot Interaction Scenario Salem et al. (2015) featured a robot acting as a home-assistant requesting the human visitor to carry out tasks that were outside of the social norm. People would comply with the robot's instructions to, water a plant with orange juice, throw away letters and use a password to login to a laptop to view and disclose confidential information. This implies that the level of trust and cooperation are high in a home scenario. When comparing this to a work scenario involving both human and robot, Desai et al. (2012) found that if there was error in the robot's performance, the perceived level of trust and collaboration would fall.

Robinette, Li, Allen, Howard, and Wagner (2016) carried out a study on an artificial emergency evacuation caused by filling a room with smoke and sounding a smoke alarm. They found that despite directing participants to evacuate to an area that was not safe, the robot's instructions were trusted and followed. The only exception to this was when participants witnessed faults during an initial guided tour given by the robot. However, the occurrence of this was higher than expected. This evidence suggests significant "over trust" in robots during emergency scenarios. Finally, the last of these scenarios, explored compliance with a robot guiding people out of a simulated maze under a time constraint (Robinette, Howard, & Wagner, 2017), either by being too slow or failing entirely, had a negative impact on compliance with the robot. The authors also note that, their scenario, although set in a natural environment, was still part of an experiment, and thus may not have invoked the same reaction as a real-life emergency scenario and that this should be considered when evaluating the results of this experiment.

1.2 Measuring Trust in Human-Robot Interaction

Across HRI research, different methods are used to measure trust, including both subjective (generally, in the form of questionnaires) and behavioural measures. Table 1 summaries the techniques used in 11 studies of trust in HRI that we have identified in the literature. It appears that the field is still largely dominated by post-hoc questionnaires (Sarkar et al., 2017; Lucas et al., 2018; Wiegmann et al., 2001; Mirnig et al., 2017; Hamacher et al., 2016; Desai et al., 2012; Salem et al., 2013), even though they are prone to post-hoc reconstruction, and raise concern regarding the actual ascription of trust (is the participant rating his/her trust in the robot or in the researcher who programmed the robot?) Interestingly, no unique validated scale exists to assess trust in the HRI domain, and people have mostly relied on study-specific questions.

Reflecting on the use of *post-hoc* questionnaires, Hancock et al. (2011) also draw awareness to the fact that such a methodology only allows to witness a singular moment of trust as opposed to an ongoing development of trust, limiting our understanding of the dynamics of trust building.

Open-ended post-session interviews are also used to assess trust. For instance, Parasuraman and Miller (2004) interviewed participants to evaluate effects of etiquette and reliability on users' rated trust in an automated system.

In contrast, behaviour-based objective measures are indirect measures of trust, but are typically less subject to post-hoc reconstruction and rationalisation. Compliance tasks (where the human is asked by the robot to perform a sequence of actions more and more committing and/or nonsensical) are the most common technique (Salem et al., 2015; Robinette et al., 2016, 2017). Willingness to cooperate is a measure from J. J. Lee,

Table 1: Overview of techniques and environments in which trust has been assessed

	Subjective measurements		Behavioural measurements		
	Post-hoc questionnaires	Interview	Compliance	Response time	Proxemics
Laboratory-based	(Sarkar et al., 2017), (Lucas et al., 2018), (Wiegmann et al., 2001), (Mirnig et al., 2017), (Hamacher et al., 2016), (Desai et al., 2012) , ours		(Robinette et al., 2017)	(Wiegmann et al., 2001)	ours
Hybrid (e.g., experimental studio)	(Salem et al., 2013)	(Parasuraman & Miller, 2004)	(Salem et al., 2015)		
Natural environment			(Robinette et al., 2016)		

Knox, Wormwood, Breazeal, and Desteno (2013), combined with the concept from Wilson, Straus, and McEvily (2006) stating that cooperation is

a "behavioural outcome of trust". Robinette et al. (2016) used an additional question post experiment questionnaire to investigate whether a participant's cooperation with the robot was due to trusting the robot guide. Response times have been used in (Wiegmann et al., 2001) where the users' agreeing with the automated aid system and their decision time are found to be related.

Questionnaires The two studies presented in this paper use either subjective measures of trust using a post-hoc questionnaire (Study 1) or behavioural measures based on proxemics (Study 2). The questionnaires used in Study 1 test several constructs:

Personality tests are used as a way to mitigate any knock-on interaction effects as a result of different personality types. The Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow, & Swann, 2003) is used to assess facets of the participants personality; namely extroversion, agree-ability, conscientiousness, emotional stability and openness to new experiences. This could have a significant impact on how a participant would rate their interaction with the robot as seen in previous research (Salem et al., 2015).

To uncover any pre-existing negative feelings towards robots, the Negative Attitude towards Robots Scale (NARS) can be utilized (Nomura & Kanda, 2003). This scale collects the participants' attitudes towards "situations of interactions with robots", "social influence of robots" and "emotions in interaction with robots" (Sarkar et al., 2017). The results of this 14 item scale are collated into three sub-scales that can be tested for correlation against final reported levels of trust to measure a possible impact.

A commonly used tool to examine a participant's experience of a human-robot interaction is the Godspeed Questionnaire. This collects the participant's perceived anthropomorphism, animation, likeability, intelligence and safety of a robot (Bartneck, Kulić, Croft, & Zoghbi, 2009).

Finally, we use additional Likert scale questions to gain targeted information and insight into a participant's impression of the robot's trustworthiness and intelligence, as in (Robinette et al., 2016).

1.3 Investigating the impact of different errors on trust

1.3.1 Research Questions

The two studies outlined within this paper share the common goal of identifying whether the nature of the errors exhibited by a faulty robot has a significant impact on participants' level of trust in the robot. Our research questions are: (1) can we robustly replicate previous results from the literature on the impact of faulty robot behaviours on trust in a short, face-to-face, lab interaction typical of a human-robot co-worker scenario? (2) if so, does a simple technical failure impact the willingness to work again with

the robot differently than a decision-level cognitive error or socio-cognitive error? and finally, (3) does the robot showing awareness of its own errors (by acknowledgement) mitigate the impact of the error on reported trust levels?

1.3.2 Hypotheses

1. *No Error vs. Erroneous Conditions*: Participants interacting with the robot that makes no errors will report a higher level of trust and willingness to work with the robotic assistant in any environment than participants interacting with the robot in both conditions where errors are made.

2. *Technical Error Condition vs. Cognitive (decision-level or socio)*: Participants experiencing robot errors will report higher levels of trust and willingness to work with the robotic assistant in any environment when the robot makes a perceived technical failure compared with a decision-level or socio-cognitive error, as a technical failure would be perceived as less serious and easier to fix.

3. *Robot Acknowledgment vs. No Robot Acknowledgement*: The acknowledgement of errors by the robot will mitigate a detrimental effect of errors on participants' reported level of trust and willingness to work with the robot, as it implies that the robot is aware of its own failure, and can possibly act on them in the future.

2 Study 1: Impact of errors and robot acknowledgement of errors on trust

The first study looks at the impact of faulty robotic behaviours on trust in a short, face-to-face interaction involving a joint assembly task typical of a human-robot co-worker scenario. The human performs the assembly of a toy, having to rely on the robot's guidance to achieve it.

2.1 Methodology

2.1.1 Experimental Procedure

The task carried out by the participants consisted of working cooperatively with the robotic assistant to complete a building task shown in Figure 1. The instructions were given to the participant by the robotic co-worker and the participant was expected to complete the building aspect of the task. The task involved building a large toy using plastic nuts and bolts. It is completed in five main stages, broken down into eleven instructions in the baseline condition, with one additional instruction required in each of the

error conditions to rectify the robot error, given by either the robot or human due to the 2×2 design of the experiment. The assembly task was designed to be easy enough to be accessible to any adult, but complex enough to be non-trivial without external guidance. In particular, many additional parts, that were not required for the assembly, were available and effectively acted as distractors.

The technical failure (*TF*) error condition involved the robot knocking items off the assembly table at the first stage after correctly pointing to two other items. Whereas, the second error condition, decision-level error (*DL*), featuring the perceived decision-level mistake, included the robot giving incorrect guidance at the very first instruction which will cause the participant not to be able to perform the last command. This would result in the participant being unable to complete the task without additional help. The baseline (no error) condition set the standard assembly instructions and level of social agency of the robotic assistant to allow for accurate comparison between the baseline and different error conditions.

We adopted a 2×2, between subject, design (Table 2). The five conditions are as follows: no error (baseline); technical failure, *TF*, with and without error acknowledgement; decision-level error, *DL*, with and without error acknowledgement.

Table 2: Condition design and sample sizes for Study 1

	Technical failure	**Decision-level**
Acknowledgement	n = 13 *(M = 6, F = 7)*	n = 15 *(M = 8, F = 7)*
No acknowledgement	n = 20 *(M = 8, F = 12)*	n = 18 *(M = 7, F = 11)*

The errors are either acknowledged by the robot in erroneous conditions with error-acknowledgement behaviour (*Ack*) or by the experimenter when the robot does not acknowledge them in erroneous conditions without error-acknowledgement behaviour (*No-Ack*). In the technical failure condition, the pieces are either collected by the participant as instructed by the robot in the *Ack* condition or by the experimenter in the *No-Ack* condition. In the decision-level error condition, the participants are provided with an additional instruction to help them rectify the error and finish the task by either the robot or the experimenter in the *Ack* and *No-Ack* conditions respectively.

Robot Control We use a TIAGo robot from Pal Robotics (Pages, Marchionni, & Ferro, 2016). The robot consists of a mobile base, a torso, an arm, a wrist, an end-effector and a head. TIAGo is 145 centimeters long

when its torso is fully extended. The arm has seven degrees of freedom ending in a gripper that enables the robot to point to the required pieces. The head features a face and has two degrees of freedom, providing pan-tilt movements to enable the robot to gaze on the pieces as it points to them. The interaction is controlled using a Wizard of Oz method (WOz). The wizard sits behind the participants, out of their field of view, as illustrated in Figure 1.

Procedure The participants first sign a consent form, then complete a pre-study questionnaire. They interact with the robotic assistant to complete the assembly task; fill the post-study questionnaire, and finally are debriefed on the experiment aims. Before leaving, the participants receive compensation for their time in the form of a voucher.

The human-robot interaction itself features a combination of verbal and physical communication. The robot provides the instruction the participant needs to complete the next step of the assembly task verbally, whilst simultaneously gazing from the participant to the objects needed and pointing with its arm. The participant were instructed to simply say 'Done!' when they were done with the current step. The role of the wizard was limited to pressing a key every time the participant had completed a step, to instruct the robot to continue to the next assembly step. This allowed for the participant to take as much time as they needed to complete each stage while avoiding possible speech recognition issues. The wizard could also get the robot to repeat the instructions for the current step if the participant expressed that he did not understood.

2.1.2 Data Collection

The pre-study questionnaire began with two demographic questions relating to the age and gender of the participants, then participants' previous experience with robots was also collected to insure a balanced distribution among the three robot conditions. The *Ten Item Personality Inventory* (TIPI) (Gosling et al., 2003) questionnaire was used to assess facets of the participants personality. In an attempt to uncover any pre-existing negative feelings towards robots, the pre-study questionnaire also included the *Negative Attitude towards Robots Scale* (NARS) (Nomura & Kanda, 2003).

The post-study questionnaire included the Godspeed questionnaire (Bartneck et al., 2009), with five sub-scales: anthropomorphism, animation, likeability, perceived intelligence and perceived safety. Participants also answered a set of 5 study-specific questions aiming at measuring trust ascription. The first four questions where 5-point Likert scales measuring how willing they would be to work with the robot again in a manufacturing environment, an office environment, a home environment or in a care centre (Trust and Willingness

to Work Scale). The fifth question asked the participants to rate the level of trust they have in the robot on a scale from 0 to 10.

2.1.3 Participants demographics

Participants were sampled from diverse backgrounds (student, university staff and local public). The final sample is made up of 100 participants (46 male, 54 female; mean age M=35.8 years, SD=13.3) after 9 were excluded due to unintentional robotic technical failures or incorrect completion of the questionnaires and in one case the participant avoiding the intentional mistake. The participants interacted with the robot for a mean interaction time M = 05:23 minutes, SD = 02:06, completing the assembly task outlined previously.

2.2 Results

Table 3: Mann–Whitney U test results for Trust and Willingness to Work Scale

	No Error vs. Faulty behaviour Hyp. 1	Technical failure vs. Decision-level error Hyp. 2	Ack. vs. No ack. Hyp. 3
Home Assistance	*U = 1202* *p = 0.54*	*U = 502* *p = 0.58*	*U = 434* *p = 0.19*
Manufacturing Environment	*U = 1158* *p = 0.77*	*U = 518* *p = 0.71*	*U = 456* *p = 0.27*
Office Assistance	*U = 1226* *p = 0.43*	*U = 496* *p = 0.51*	*U = 427* *p = 0.15*
Caring for a Family Member	*U = 1328* *p = 0.12*	*U = 624* *p = 0.29*	*U = 490* *p = 0.57*
Trust Level	*U = 1416* *p = 0.03**	*U = 508* *p = 0.63*	*U = 471* *p = 0.43*

Independent T-tests were carried out on the subscales generated from both the TIPI and NARS tools used in the pre-study questionnaire in conjunction with the data collected using the post-study Trust and Willingness to Work Scale. In summary, we only found a weak yet statistically significant correlation between subscale 2 of the NARS and the level of trust (r=-0.449, p=0.004), i.e. the more negative the participants' views of the

social influence of robots the lower the perceived level of trust. No interactions were found for the Ten Item Personality Inventory (TIPI). Finally, only one significant interaction was found with the Godspeed questionnaire: the robot in the *TF* condition is statistically more likeable than in the no-error condition (s=-2.095, p=0.046).

2.2.1 Hypothesis 1: No error vs. erroneous conditions

In order to test this hypothesis, non-parametric Mann-Whitney U tests (as the answers did not follow a normal distribution – see Figure 2) were carried out on the results of the Trust and Willingness to Work Scale between the no-error and erroneous conditions which includes both technical failure and decision-level errors.

Figure 2 shows the distribution of trust and willingness to work with the robot again in the four investigated environments for the control group (no error condition) against the technical failure and decision-level errors. The U-test values reported in Table 3 provide no statistically significant evidence to support Hypothesis 1 in the four evaluated environments. However, Hypothesis 1 is partially supported with regards to the reported trust with $U = 1416$, $p = 0.03$, and an effect size of $P(trust_{ctrl} > trust_{faulty}) = \frac{U}{n_{ctrl} \cdot n_{faulty}} = 63\%$ (probability that one random observation from trust values of the control conditions is larger than a random observation from trust values of the error condition; large effect).

2.2.2 Hypothesis 2: Technical failure vs. decision-level error conditions

Similar to our first hypothesis, the second hypothesis is also investigated by performing Mann-Whitney U tests on the same variables but between technical failure conditions with and without robot acknowledgement grouped together and decision-level error conditions with and without acknowledgement grouped together as well.

The distributions of trust and willingness to work with the robot again in the four investigated environments for the grouped technical failure error conditions and the grouped decision-level error conditions are depicted in Figure 2. The U-test values of these tests are listed in Table 3. These results show no impact of the type of the error experienced by the participant on the examined variables.

2.2.3 Hypothesis 3: Acknowledgement vs. no acknowledgement when a fault occurs

Hypothesis 3 is also tested by applying Mann-Whitney U tests on the evaluated variables between the participant groups interacting with a robot acknowledging its errors and a robot which does not acknowledge them.

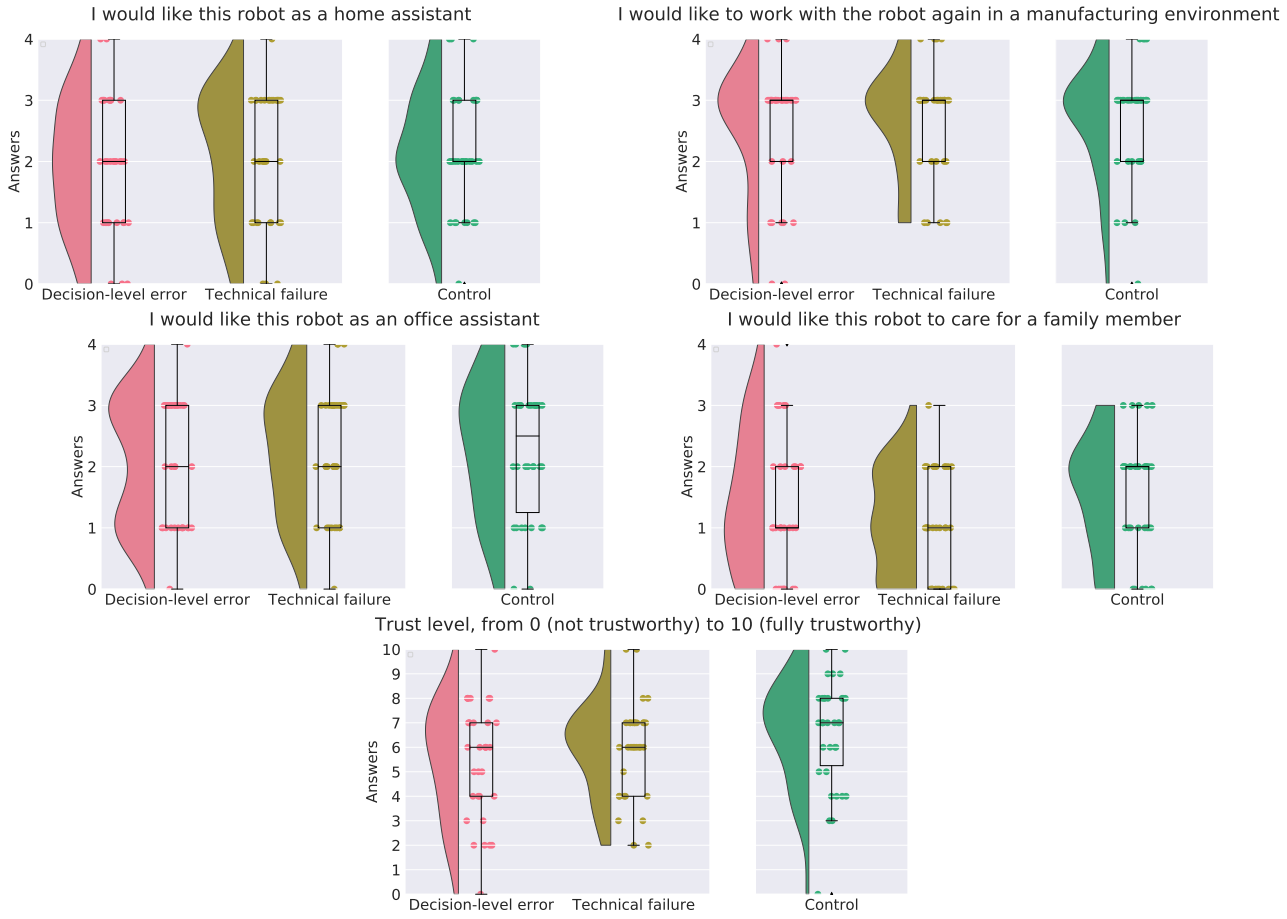

Figure 2: **Impact of error type**. Distributions of willingness to work with the robot again in the four investigated environments (0=fully disagree; 4=fully agree), as well as reported trust level, where the type of error (technical failure vs. decision-level error) is the independent variable. RainCloud plots (Allen et al., 2018) are used.

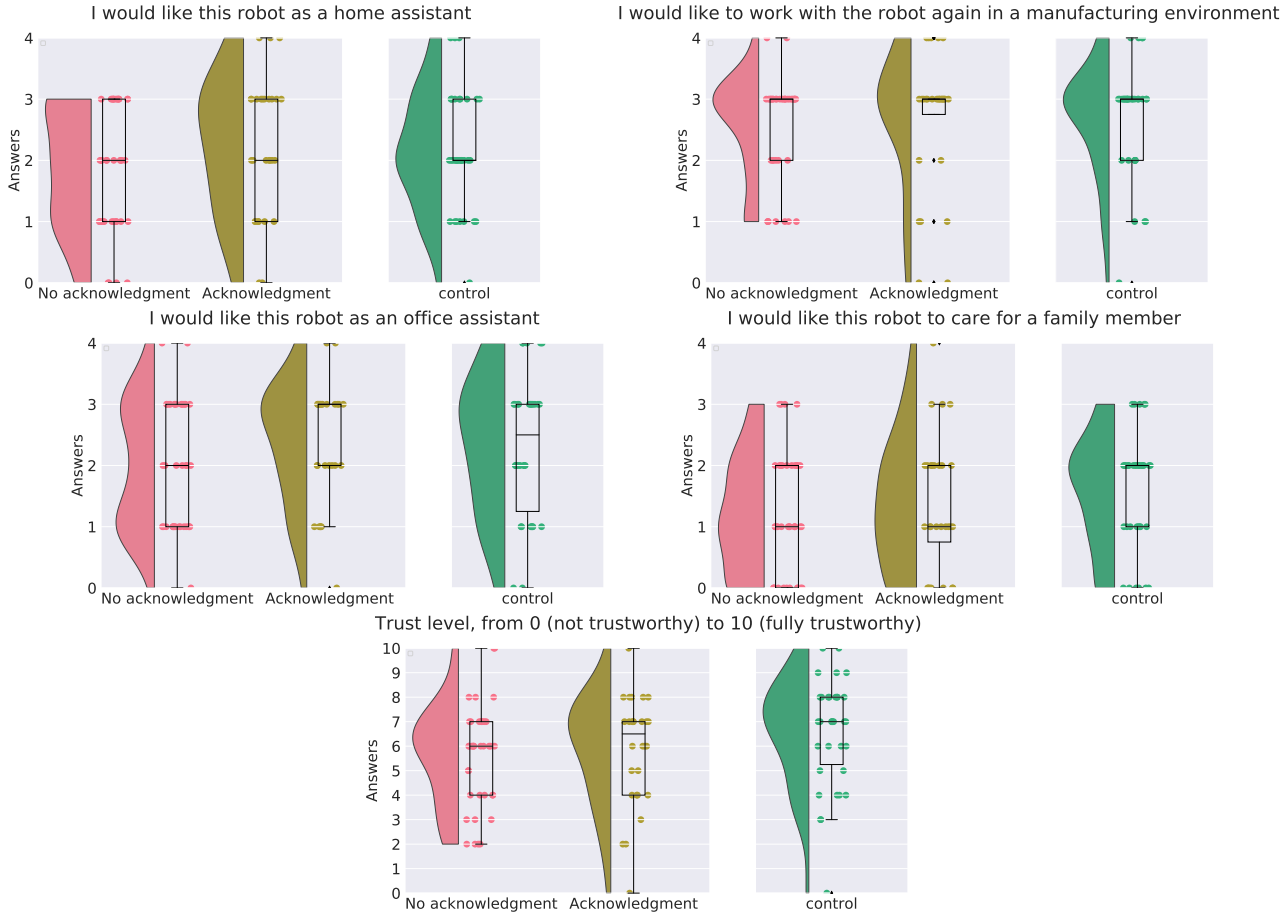

Figure 3: **Impact of error acknowledgment**. Distributions of willingness to work with the robot again in the four investigated environments (0=fully disagree; 4=fully agree), as well as reported trust level, where the acknowledgment or non-acknowledgement of error is the independent variable.

Figure 3 illustrates the distributions of the examined variables' values for a faulty robot, when it does or does not acknowledge its errors. Table 3 reports the U-test results. No significant difference in the reported trust level and the willingness to work with the robot again in the four investigated environments were found.

2.2.4 Errors and acknowledgement behaviours conditions internal interactions

For each evaluated variable, trust and willingness to work with the robot again in the four different environments, the two independent factors (error and acknowledgement behaviour) have two levels each. This yields four different combinations as illustrated in Table 4. To fully investigate all potential impacts that might have resulted from the interaction between these combinations, Mann-Whitney tests were performed on the examined variables. The tests showed no statistically significant impact of any combination of the independent factors' levels on trust and willingness to work with the robot again in the four different environments[1].

Table 4: The four combinations of the different levels of the two independent factors (error and acknowledgement behaviour)

	Technical Failure	Decision-level
Acknowledgement	*TF* & *Ack*	*DL* & *Ack*
No Acknowledgement	*TF* & *No Ack*	*DL* & *No Ack*

3 Study 2: Impact of Errors on Proxemics

Like the first study, the second study looks at the impact of error types on trust levels. However, this study (performed independently of the first one, and led by a different researcher) uses behavioural measurements (based on proxemics) to assess trust.

[1]The values of the 20 tests (four combinations with five variables each, trust and willingness to work with the robot again in the four different environments) are provided online as indicated in Section 6.

3.1 Methodology

3.1.1 Experiment Procedure

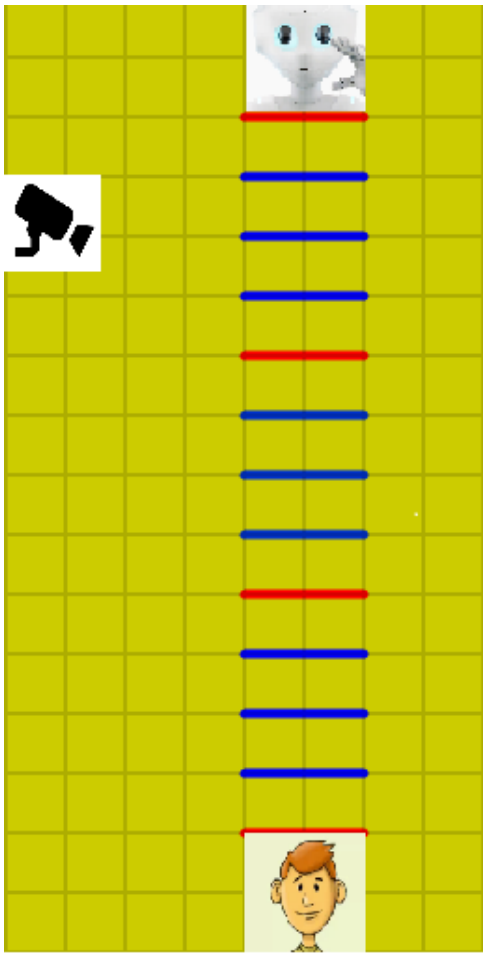

Figure 4: Experimental setup for Study 2. Participants are stood in front of the robot; each line on the floor is marked at 25 cm, so we can measure the stop distances between the robot and the participant. The participant stands initially 3m from the robot.

Each participant had to perform three tasks, for which so-called 'stop distances' were measured (Figure 4). These distances were:

- *Human stop distance*: participants were instructed to walk towards the robot and stop whenever they felt they did not want to come any closer to the robot.
- *Back off distance*: participant would stand face-to-face with the robot

as close as possible and then were asked to slowly walk backwards and stop whenever they felt comfortable again.

- *Robot stop distance*: the robot would start at a distance of 3 meters and approach the participant. Whenever the participant started to feel uncomfortable and wished the robot would not come any closer, they would say 'Stop' and the robot would stop.
- *Stop distance difference*: In order to get an idea about the relation between robot stop distance and human stop distance, this measurement was recorded as well. This is nothing more but the robot stop distance subtracted from the human stop distance.

The order of the tasks (human-initiated or robot-initiated) was counterbalanced across participants. The robot used for this experiment is Pepper from Soft Bank Robotics. Participants were randomly assigned one of three conditions. In two of three conditions, the robot shows faulty behaviour during the introduction, before the tasks mentioned above were performed. These conditions are:

- No error: the robot approaches the participants at a speed of 1.8 km/h without saying anything.
- Technical error: the robot 'accidentally' knocks over a pile of items beside it while waking up from its default state (Figure 5). The items are placed in such a way that the collision was not expected.
- Socio-cognitive error: the robot incorrectly recognizes the experimenter's gender during the introduction, where the experimenter mentions the robot is capable of doing so.

Observing Pepper make a socio-cognitive error (gender confusion) is hypothesized to negatively impact the robot's perceived intelligence rating and the approach distance. This is supported by Salem et al. (2015) who found that a robot's faulty behaviour caused a change in the robot's perception. Observing Pepper make a technical error will impact the approach distance as well as its perceived intelligence and perceived safety.

In the error conditions, the robot does not acknowledge its mistake.

3.1.2 Data Collection

The experiment started with collecting consent and demographics (including previous experience with robots). Similar to the previous study, TIPI questionnaire were used to investigate whether certain personality traits affected the results. During the study, the different stop distances (dependent variables) mentioned before were measured. Post-study questionnaires involved the Godspeed questionnaire, together with questions regarding the participant's current mood and their perceived safety during the experiment.

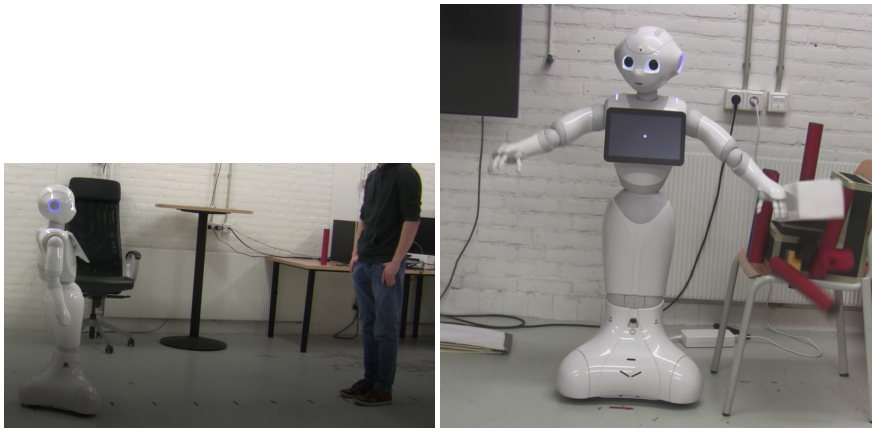

Figure 5: Left, Pepper during its approach of the participant; right, Pepper knocking over items when stretching.

3.1.3 Participant Demographics

In total 60 adults (29 male, 31 female; age $M = 33.8$ years, $SD = 15.9$; min age = 18, max age = 75) from different backgrounds (students, working public, retirees) took part in the experiment. The majority (93%) of these participants were Dutch (other nationalities include German, Spanish and Bulgarian). All participants completed the experiment. Participants were randomly assigned to either the control condition ($n = 20$, 13 female, 7 male), the *Technical error* condition ($n = 20$, 7 female, 13 male), or the *Social error* condition ($n = 20$, 11 female, 9 male). Participants had no to little experience with robots ($M = 1.52$, $SD = 1.03$ on a scale from 1 (no experience) to 5 (very experienced)).

3.2 Results

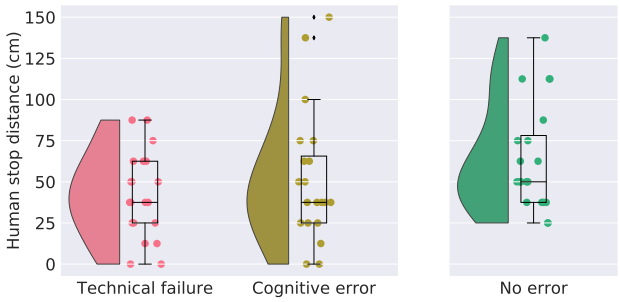

Figure 6: Distance (in cm) at which participants stop getting closer to the robot. The control condition is plotted on the right-hand side.

3.2.1 Faulty behaviour vs baseline

A two-way MANOVA was performed to look at possible interactions. The independent variables were the condition (error or baseline) and the approach order (human first or robot first), while the recorded approach distances and the stop distance difference were the dependent variables. The results showed a significant difference for the human stop distance, with Pepper being approached closer in the error condition compared to the baseline condition ($p = 0.015$). This means that the participant approached closer when a technical error was observed compared to no error being observed. The stop distance difference differed significantly ($p = 0.001$), as the robot was told to stop earlier after observing a technical error compared to not observing an error beforehand. This same difference was found between social error and baseline ($p = 0.001$).

To investigate whether there was a difference in perception between the error condition and the baseline, we also performed Mann-Whitney U tests using the questionnaires. The independent variables were the conditions (error or baseline) and the dependent variables were the median scores on the Godspeed questionnaire. A significant difference was found for anthropomorphism between the technical error and baseline ($U = 119.5$, $p = 0.025$), where the robot was scored as less anthropomorphic after making a technical error. Significant differences were also found for anthropomorphism ($U = 114.5$, $p = 0.015$) and animacy ($U = 105$, $p = 0.007$) between the social error condition and the baseline, where both factors got a lower score after making a social error. No other significant differences were found between the error conditions and baseline regarding the perception of the robot.

3.2.2 Technical vs Socio-cognitive error

A two-way MANOVA was run to investigate whether a different type of error had an influence on the different approach distances (robot stop distance, human stop distance, back off distance and stop distance difference). The independent variables were the two error conditions and the two different orders of approach while the dependent variables were the three measured distances and the stop distance difference. As a representative illustration, Figure 6 shows the results for the 'human stop distance' metric. The analysis showed no significant difference on the approach distances between the technical error and social error:

- robot stop distance: ($p = 0.904$)
- human stop distance: ($p = 0.352$)
- back off distance: ($p = 0.558$)
- stop distance difference: ($p = 0.202$)

The order of approach had a significant influence on the robot stop distance ($p = 0.006$) and the stop distance difference ($p = 0.002$). The order did not have a significant influence on back off distance ($p = 0.639$) and the human stop distance ($p = 907$). No interaction effects between type of error and approach order were found. These results indicate that when the participant is the first to approach the robot, then the stop distance becomes smaller. When the robot is the first to approach, then the delta between the robot and the human stop differences becomes larger, with the robot stop distance being larger than the human stop distance.

Mann-Whitney U tests were performed to investigate whether there was a difference in how the robot was perceived after witnessing the robot make either a technical or a social error. The results showed that there was no significant difference between the two error conditions for how the robot was perceived. For anthropomorphism the results were ($U = 197.5$, $p = 0.944$), for animacy ($U = 188.5$, $p = 0.751$), for likeability ($U = 147.5$, $p = 0.116$), for perceived intelligence ($U = 192$, $p = 0.814$) and for perceived safety ($U = 198$, $p = 0.952$). This means that there is no difference in the type of error as far as perception of the robot is concerned in the five factors of the Godspeed questionnaire. When looking for any correlation between the various stop distances and the five personality traits from the TIPI questionnaire, none were found, which means that there seems to be no clear correlation between any of the personality traits and the distance people stopped approaching or told the robot to stop. This means the TIPI results can not be used to predict the distances.

4 Discussion

We have presented two studies investigating the impact of different types of error on ascribed levels of trust, totalling the inclusion of 160 participants. In both studies, we found a general impact of errors on reported and observed levels of trust. These results are, however, weak: in Study 1, only the *Trust* ratings did change significantly, but none of the four other questions related to the willingness to use the robot again in specific environments did. In Study 2, the difference between the no-error condition and the faulty condition was counter intuitive, as participants came actually significantly closer to the faulty robot. Despite the results appearing weak, this is often the outcome when studying trust in HRI due to a number of later discussed confounds. To example this, Mirnig et al. (2017) also found erroneous robot behaviours resulted in no impact on anthropomorphism and perceived intelligence. The same study reported a significant increase in likeability in the error condition, a possible attribute of participant novelty to interacting with a robot, resulting in increased patience levels (Mirnig et al., 2017).

Thus, no definitive conclusion can be reached regarding our main re-

search question, the impact of error types on trust: in Study 1, we compared a technical failure to a higher-level cognitive failure (wrong decision) with no significant impact, and in Study 2, we compared a technical failure to a socio-cognitive error (gender confusion) with, again, no significant difference.

Three main explanations for this lack of conclusive results can be considered: *(1)* the type of errors has indeed little impact on the perceived robot trustworthiness; *(2)* our tasks were not suitable to effectively measure (possibly subtle) differences in trust ascription between our conditions; or *(3)* the low ecological validity of the experimental environment (short interactions in a laboratory setting) did overshadow any effect (measure sensitivity issue). The latter two confounds are plausible, and we discuss them hereafter.

4.1 Potential confounds

Regarding the choice of task, the low severity of the tasks in both studies may have led to a limited impact on the participants' feelings after taking part in the study: in Study 1 in particular, the participants were building a children's toy, with no time constraint or implications of incorrect assembly (beyond having to backtrack a few simple steps). The effects of the robot performing an error were limited to mere annoyance. This could have been compounded by the fact that the interaction with a robot was for the majority of participants novel and potentially exciting, meaning the participants enjoyed experiencing an interaction with a robot in any case, which then overshadowed the consequences of the robot's error.

Another potential confound relates to the appearance of the robots chosen for these studies – with long-established models like the Uncanny Valley postulating that human-looking robots might be found to be unnerving by humans. Gray and Wegner (2012) investigated the reasoning behind this theory, suggesting that a human-like appearance might lead humans to project a sense of mind onto a robot. This study found that people are not only unnerved by a robot with a humanoid appearance, but also a robot having a sense of experience and this same sense lacking in fellow humans. Goetz et al. (2003) found that when using robots that appear to be male, people would prefer a machine-like robot when performing a realistic (e.g. Office Clerk or Hospital Message and Food Carrier) or conventional (e.g. Soldier or Security) job role. The human-like "male" robot was only preferred in artistic (e.g. Actor) or social (e.g. Tour Guide) roles. The researchers found more significant results when testing a "female" robot. A machine-like robot was preferred for investigative roles (e.g. Lab Assistant) and also realistic job roles, but a human-like robot in all other job areas: artistic, enterprise (e.g. Sales Representative), conventional and social. TIAGo is a machine-like "male" robot, so it was chosen for the assembly task, which would most likely be classified by a naïve user as an investigative or con-

ventional role. Study 2, however, used Pepper, compared to TIAGo a more human-like, "female", robot, yet showed no difference in the ascription of trust.

Regarding explanation *(3)* (low ecological validity), experiments carried out in a lab setting are likely to be perceived as artificial and controlled (Baxter, Kennedy, E., Lemaignan, & Belpaeme, 2016), and as such, generally safe. This in turn reduces the potential impact of the introduced errors, as no severe consequences are to be expected.

Also, the participants' reported level of trust may possibly be subconsciously attributed to the experimenter and not the robot. This is a knock-on effect of not carrying out a study 'in the wild', and therefore having low levels of realism and low ecological validity.

Besides, as participation was voluntary (and the compensation small), our experimental population must have had an intrinsic interest for robots, that would skew the attitude towards robots toward positive feelings and a stronger inclination to trust the robot.

Finally the participants' reported level of trust and intelligence might possibly have been subconsciously attributed to robots in general rather than to the specific robots that were used for these two studies. This could have caused the invariance in the reported levels of trust and intelligence between the control and erroneous conditions and among the erroneous conditions.

4.2 A lack of negative results?

In light of these several potential confounds, one might rightfully question how suitable a laboratory environment is for the study of trust. We acknowledge that even broader discussions on the limits of lab environments to conduct HRI studies have already been made, for instance (Baxter et al., 2016). Yet, as we show in Table 1, most of the existing literature on trust in HRI reports on studies performed in lab environments, often using subjective measures (post-hoc questionnaires) that are subject to a lot of hard-to-control interpersonal noise. Our two studies show that, even with reasonable sample sizes (100 for Study 1, 60 for Study 2) and using both subjective and objective measures, we find weak and/or inconsistent results. As a result of the replicability crisis that has been much discussed over the past few years, we can only recommend for more replication studies, and for our community to embrace the publication of negative results (through pre-registered studies, for instance), in order to build a better understanding of the experimental 'degrees of freedom' that are available to us when investigating trust.

5 Conclusion

This article investigated the impact of different types of errors on participants' reported levels of trust in a robotic assistant. The first study (a robot-guided assembly task) did evidence some effects of errors on trust: while we found a significantly lower ascription of trust on the faulty robot compared to the control group (in particular when the robot does not acknowledge its errors), no effects of the type of errors (mechanical vs. decision-level) on trust were found, and neither errors had impact on the willingness to use the faulty robot again in a different environment at a later point.

Using proxemics instead of questionnaires to measure trust, our second study found broadly similar results, with an effect of errors on the willingness to move closer to the robot (however, opposite to the intuition: people would get closer to the faulty robot), but no significant impact of the error type on the participants' behaviour.

In order to further investigate the lack of a significant difference between types of error, we contrasted as well a robot acknowledging errors (and henceforth, demonstrating an awareness and understanding of the situation) with a robot that did not demonstrate such awareness of its own errors. No significant difference between these two conditions were found.

Even though *some* level of trust manipulation was successfully performed in our lab environment, more subtle effects were not clearly evidenced, and we attribute this lack of results to the lab environments not generally providing sufficient sensitivity to measure complex social constructs like trust.

As such, our conclusion is that neither of our two studies provide conclusive evidence regarding the impact of the type of errors on the resulting evoked trust in robots, and that furthermore, the robot acknowledging or not its errors does not automatically lead to significant changes in perception.

6 Resources for Replication

Following recommendations by Baxter et al. (2016), we briefly outline hereafter the details required to replicate our findings.

Study The experimental protocol has been provided in the text. Exact robot dialogues, detailed questionnaires, as well as the open-source code for the wizarding interface are available online: `https://git.brl.ac.uk/ra3-flook/Trust-vs-Errors`.

Data analysis The full recorded experimental datasets, for both studies, as well as the data analysis scripts allowing for reproduction of the results and plots presented in the paper (using the Python *pandas* library) are open and available online (`https://git.brl.ac.uk/ra3-flook/Trust-vs`

`-Errors`). The script includes all pair-wise group comparisons across all conditions.

7 Acknowledgements

Part of the work has been funded through the UK EPSRC Riveras project, grant EP/J01205X/1.

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. (2018, August). Raincloud plots: a multi-platform tool for robust data visualization. *PeerJ Preprints*, *6*, e27137v1. Retrieved from `https://doi.org/10.7287/peerj.preprints.27137v1` doi: 10.7287/peerj.preprints.27137v1

Barber, B. (1983). The logic and limits of trust.

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. doi: 10.1007/s12369-008-0001-3

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The influence of people's culture and prior experiences with aibo on their attitude towards robots. *Ai & Society*, *21*(1-2), 217–230.

Baxter, P., Kennedy, J., E., S., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of hri to methodology and reporting recommendations. In *Proceedings of the 2016 acm/ieee human-robot interaction conference (alt.hri)*. doi: 10.1109/HRI.2016.7451777

Bickmore, T., Pfeifer, L., Schulman, D., Perera, S., Senanayake, C., & Nazmi, I. (2008). Public displays of affect: Deploying relational agents in public spaces. In *Proceedings of chi'08* (pp. 3297–3302). doi: 10.1145/1358628.1358847

Breazeal, C., Kidd, C., Thomaz, A., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of the ieee international conference on intelligent robots and systems* (pp. 708–713). doi: 10.1109/IROS.2005.1545011

Corritore, C., Kracher, B., & Wiedenbeck, S. (2003). Online trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, *58*(6), 737–758.

Dautenhahn, K., Woods, S., Kaouri, C., Walters, M., Koay, K., & Werry, I. (2005). What is a robot companion – friend, assistant or butler. In *Proceedings of the ieee international conference on intelligent systems and robots* (pp. 1192–1197). doi: 10.1109/IROS.2005.1545189

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, Bruggeman, S., ... Yanco, H. (2012). Effects of changing reliability on trust of robot systems. In *Proceedings of the acm/ieee conference on human-robot interaction* (pp. 73–80). doi: 10.1145/2157689.2157702

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behaviour to tasks to improve human-robot interaction. In *Proceedings of ieee roman international workshop on robot and human interactive communication* (pp. 55–60).

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big five personality domains. *Journal of Research in Personality*, *37*, 504–528. doi: 10.1016/S0092-6566(03)00046-1

Gray, K., & Wegner, D. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. doi: 10.1016/j.cognition.2012.06.007

Guznov, S., Lyons, J., Nelson, A., & Woolley, M. (2016). The effects of automation error types on operators' trust and reliance. In S. Lackey & R. Shumaker (Eds.), *Virtual, augmented and mixed reality* (pp. 116–124). Cham: Springer International Publishing.

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). doi: 10.1109/ROMAN.2016.7745163

Hancock, P., Billings, D., Schaefer, K., Chen, J., de Visser, E., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. doi: 10.1177/0018720811417254

Iwamura, Y., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2011). Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets. In *Proceedings of the acm/ieee international conference on human-robot interaction* (pp. 449–456). doi: 10.1145/1957656.1957816

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

Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., & Desteno, D. (2013). Computationally modelling interpersonal trust. *Frontiers in Psychology*, *4*(893). doi: 10.3389/fpsyg.2013.00893

Lee, M., 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). doi: 10.1145/1718918.1718927

Lemaignan, S., Fink, J., & Dillenbourg, P. (2014). The dynamics of anthropomorphism in robotics. In *in proceedings of the interna-*

tional conference on human-robot interaction (pp. 226–227). doi: 10.1145/2559636.2559814

Lucas, G., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., ... Leuski, A. (2018). Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (pp. 344–351). doi: 10.1145/3171221.3171258

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

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. In *Frontiers in robotics and ai.* doi: 10.3389/frobt.2017.00021

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, *21*(4-5), 203–211.

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Muir, B., & Moray, N. (1996). Trust in automation: Part II. "experimental studies of trust and human intervention in a process control simulation.". In *Ergonomics* (pp. 429–460). doi: 10.1080/00140139608964474

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.

Nass, C., & Lee, K. (2000). Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *Proceedings of chi'00* (p. 329 - 336). doi: 10.1145/332040.332452

Nomura, T., & Kanda, T. (2003, Nov). On proposing the concept of robot anxiety and considering measurement of it. In *The 12th ieee international workshop on robot and human interactive communication, 2003. proceedings. roman 2003.* (p. 373-378). doi: 10.1109/ROMAN.2003.1251874

Pages, J., Marchionni, L., & Ferro, F. (2016). Tiago: the modular robot that adapts to different research needs. In *International workshop on robot modularity, iros.*

Parasuraman, R., & Miller, C. (2004). Trust and etiquette in high-criticality automated systems. *Communication of the ACM*, *47*(4), 51–55. doi: 10.1145/975817.975844

Ray, C., Mondada, F., & Siegwart, R. (2008). What do people expect from robots? In *Proceedings of the ieee/rsj 2008 international conference on intelligent robots and systems* (pp. 3816–3821). doi: 10.1109/IROS.2008.4650714

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. doi: 10 .1109/THMS.2017.2648849

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *The eleventh acm/ieee international conference on human robot interaction* (pp. 101–108). doi: 10.1109/HRI.2016.7451740

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joulbin, F. (2013). To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likeability. *International Journal of Social Robotics*, *5*, 313–323. doi: 10.1007/s12369-013-0196-9

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). doi: 10.1145/2696454.2696497

Sarkar, S., Araiza-Illan, D., & Eder, K. (2017). Effects of faults, experience, and personality on trust in a robot co-worker. *arXiv preprint arXiv:1703.02335*.

Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2006). Interactive humanoid robots for a science museum. In *Proceedings of the 1st acm sigchi/sigart conference on human-robot interaction* (pp. 305–312). doi: 10.1109/MIS.2007.37

Sidner, C., Lee, C., & Lesh, N. (2003). Engagement rules for human-robot collaborative interactions. In *Proceedings of the ieee international conference on systems man and cybernetics* (pp. 3957–3962). doi: 10.1109/ICSMC.2003.1244506

Thrun, S., Schulte, J., & Rosenburg, C. (2000). Interaction with mobile robots in public places. *IEEE Intelligent Systems*, 7–11.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. In *Theoretical issues in ergonomic science* (p. 352–367). doi: 10.1080/14639220110110306

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

Wortham, R., Theodorou, A., & Bryson, J. (2016, 04). Robot transparency, trust and utility..